Management of Data Quality when Integrating Data with Known Provenance

Maria del Pilar Angeles

Submitted for the degree of Doctor of Philosophy

Heriot-Watt University
School of Mathematical and Computing Sciences
Edinburgh, UK

April 2007

Abstract

Users querying a Database System get returned a set of data with no indication of the qualitative value of that data, so the presumptions have to be that data is perfect, original and atomic. Existing database systems are based on these Presumptions of Perfection, Primary Authorship, and Atomicity. However, we know these presumptions are invalid through a considerable body of existing research. Therefore, this research seeks to challenge these presumptions.

Our research hypothesis was to identify usable quality criteria to meassure and assess data quality of data sources at multiple leveles of granularity, and derived data. These can be enhanced by the use of provenance, and the qualitative measures can be used to derive ranking of data sources based on the specification of context by the users utilising this known criteria, all within heterogeneous multi-database environment.

We propose a Data Quality Manager (DQM) composed by a generic Data Quality Reference Model, a Measurement Model, and an Assessment Model. The Reference Model provides a new general structured classification of existing data quality properties considering different user perspectives. The Measurement Model extends existing metrics for the estimation of data quality at database, relation, tuple and attribute levels of granularity, which is novel. The assessment of derived data through the use of data provenance is novel. We identify a new assessment-oriented classification based on the levels of granularity assessed. The facility to permit users to define query context in terms of quality criteria, quality prorities, and levels of granularity is also novel.

We implemented the DQM as a proof of concept of our hypothesis and demonstrate that the prototype performs approapriately according to specific requirements and can provide qualitative information, which varies according to the context.

Dedication

This thesis is dedicated to my beloved girls, Pilar and Victoria.

Acknowledgements

I am very grateful to my academic supervisors Prof. Lachlan MacKinnon, and David Marwick for their invaluable advice, support, comments, contributions and friendship.

I sincerely acknowledge guidance of previous researchers who have influenced my work.

I am also indebted to my mother and my husband who had to endure the vicissitudes of my writing up, providing me love, and inspiration to finish this step in my life.

I would like to thank my friends for their encouragement they gave me throughout my studies.

The Ph.D. and this thesis were funded by CONACYT. This funding is greatly appreciated.

My special gratitude to the Heriot-Watt University for all the facilities and services provided.

ACADEMIC REGISTRY
Research Thesis Submission

Name:	MARIA DEL PILAR ANGELES		
School/PGI:	SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES		
Version: *(i.e. First, Resubmission, Final)*	FINAL	Degree Sought:	PhD

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate:		Date:	24/04/07

Submission

Submitted By *(name in capitals)*:	MARIA DEL PILAR ANGELES
Signature of Individual Submitting:	
Date Submitted:	24/04/07.

For Completion in Academic Registry

Received in the Academic Registry by *(name in capitals)*:	K WALLACE		
Method of Submission *(Handed in to Academic Registry; posted through internal/external mail)*:	BY HAND		
Signature:	K Wallace	Date:	24/4/07

CONTENTS

Table of Tables

Table of Figures

Glossary

ACID	Atomicity, Consistency, Isolation, and Durability
AIMQ	AIM Quality Methodology
CIS	Collaborative Information Systems
CRM	Customer Relationship Management
DaQuinCIS	Data Quality in Cooperative Information Systems
DBMS	Database Management Systems
DEA	Data Envelopment Analysis
DQ	Data Quality
DQAM	Data Quality Assessment Model
DQM	Data Quality Manager
DQMM	Data Quality Measurement Model
DQRM	Data Quality Reference Model
EAI	Enterprise Application Integration
EII	Executive Information Integration
ERP	Enterprise Resource Planning
IQ	Information Quality
IQA	Information Quality Assessment
IS	Information System
ISE	Information Search Environment
IT	Information Technology
MADM	Multi Attribute Decision Making
MIT	Massachusetts Institute of Technology
NISS	National Institute of Statistical Sciences
PSP/IQ	Product and Service Performance/Information Quality
QCA	Query Correspondence Assertions
RW	Real World
SAW	Simple Additive Weighting
TDQM	Total Data Quality Management
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution

Publications

Some of the material presented in this thesis has already appeared in the following conference papers:

- Angeles, P., MacKinnon, L., "Detection and Resolution of Data Inconsistencies, and Data Integration using Information Quality Criteria", in *Quality: The bridge to the future in Quality Information and Communication Technologies (ICT)*, October 2004, Porto, Portugal, pp. 87-94.

- Angeles, P., MacKinnon, L., "Tracking Data Provenance with a shared metadata", In *Postgraduate Research Conference in Electronics, Photonics, Communications and Networks, and Computing Science*, March 2005, Lancaster, U.K., pp.120-121.

- Angeles, P., MacKinnon L., "Quality Measurement and Assessment Models including Data Provenance to grade Data Sources", In *International Conference on Computer Science and Information Systems*, June 2005, Athens, Greece, pp. 101-118.

Further papers based on the material presented in this thesis are currently in preparation for journal submission.

Chapter 1 Introduction

1.1 Introduction

Users querying a Database System get returned a set of data with no indication of the qualitative information of that data, so the presumption has to be that data is 100% perfect. Most of the existing database systems are based on this "Presumption of Perfection". This is inappropriate because we know that not all data in a database are necessarily perfect, *"Real world data is dirty"* [Hernandez98].

When data that is coming from a data source without any other information, users have to assume that the source is a primary data source and that the data value is atomic. However, given the fact of the explosion of online databases created dynamically it is highly unlikely that in any data source all the data would have been primarily authored for that database or that all the values in there are necessarily atomic. These presumptions, the "Presumption of Primary Authorship", and the "Presumption of Atomicity" exist because we have no other way of dealing with data coming from data sources in the existing situation.

Typically, different systems can give you conflicting answers to the same question *"Good answers from bad data"* [Kno95]. Extensional data inconsistencies are derived from the integration of independent, distributed data sources. Syntactic and semantic inconsistencies are revealed from the differences between schema, representation, and data values among the participant data sources [Motro98].

We therefore need to be able to identify mechanisms, by which we can determine the original source of data and the original atomic values, which could be used for derived data quality estimation.

We present a Data Quality Manager that provides data consumers with qualitative information of data sources, and derived data within heterogeneous multi-data source environment. Therefore, such qualitative information is given at multiple levels of granularity. This qualitative information could help users when facing extensional data inconsistencies.

Rather than the presumption of perfection, we have identified from existing research a set of quality criteria to provide accurate measures of quality for each item of data returned from the data source. Rather than the presumption of primary authorship, we used data provenance as a mechanism to identify the origination of data within data sources, and cascade qualitative information along the ancestor trail. Rather than the presumption of atomicity, we can now either assure the user the data is atomic or identify that it is a composed data and identify the atomic values from it was generated.

1.2 Problem Definition

In an effort to control data, a number of approaches for data integration have been developed such as Enterprise Application Integration (EAI) [Linthicum99], and Enterprise Information Integration (EII) [Fensel05]. Even, taking an application oriented approach by installing Customer Relationship Management (CRM) [Rajola03] or Enterprise Resource Planning (ERP) systems to aggregate and manage enterprise data [Shtub99]. However, *"Each approach focuses on the movement or synchronization of data not the quality of data itself"* [Loshin05b], and *"Developers have to create transformation programs manually to deal with heterogeneous data sources in a company"* [Fensel05]. Therefore, the direct consequences are poor data quality and extensional data inconsistencies [Motro93].

In addition to these enterprise approaches, a considerable body of research over a period of 35 years in heterogeneous databases has also considered and attempted to resolve the issues of extensional inconsistencies [Sheth90], [Sheth92], [MacKinnon98], [ElKathib00].

Integration of schemas on existing databases into a global unified schema, is an approach developed over 20 years ago [Batini86]. However, few approaches have focused on the data value level [Anokhin01]. Data is dependent not only on the data design such as establishing correct constraints validation [Wand96], but also on how data was generated e.g. data fusion, data transformation, or data replication. Therefore, data quality degradation is a consequence. *"Although the quality and integrity of the data at each individual component database can be high, concerning the integrated, global view of data, quality and integrity can be poor"* [Gertz98b].

In the background section of this thesis, we consider previous approaches to deal with poor data quality and data inconsistencies during the data integration processes. After this review, it can be seen that users are unable to trust one data source more than other. This situation is mainly because users have not been considered by establishing which quality criteria shall be used to determine the most convenient data for them. Allowing the user the opportunity to identify the criteria that are the most important to them profiles the user in that context.

In summary, user priorities, generation of data, and data quality differences among the participating sources have not been fully addressed for coping with extensional data inconsistencies during the process of information integration.

1.3 Motivations for Research

The prime motivation for the research is that when users query a database system, they get returned a set of data which is inherently presented as perfect, original, and atomic. Users have no information by which to judge its quality and whether data comes from a number of data sources or by a transformation function.

We therefore present a Data Quality Manager (DQM) for the measurement, and assessment of the quality of derived data, and data at different levels of granularity within a multi data source environment based on a set of quality criteria, and data provenance to return qualitative information to the users.

It would be reasonable to assume that by providing data consumers with qualitative information, they could utilise it to deal with extensional data inconsistencies.

When the research began we were focused on the consideration of the issues of data quality in providing qualitative information to users querying databases. Through the process of the research these were clarified by considering data provenance as described in Chapter 4 which gives us the following research hypothesis.

1.4 Research Hypothesis

"It is possible to identify usable data quality criteria to measure and assess data quality of derived data, and data at multiple levels of granularity. These can be enhanced by the use of provenance, and the qualitative measures can be used to derive a ranking of

data sources based on the specification of context by the users in a heterogeneous multi-database environment".

1.5 Thesis Objectives

The objectives of the research to fulfil the research hypothesis are as follows:

1. The identification from existing research of a set of data properties as quality indicators to measure, assess and rank data sources, namely the Data Quality Reference Model (DQRM) referred to in Section 3.4.

2. The identification of existing metrics to be used as data quality measurement instruments at database, relation, and attribute levels of granularity of primary data sources, namely the Data Quality Measurement Model (DQMM), detailed in Section 3.5.

3. The identification of existing processes required to represent, to interpret, and to assess data quality, namely the Data Quality Assessment Model, (DQAM), see Section 3.5.

4. The implementation of a data provenance algorithm to help assessment processes for derived data sources, referred to in Chapter 4.

5. The identification and implementation of existing Multi-Attribute Decision Making methods to provide an overall quality score to rank data sources, covered in Chapter 5.

6. The design and development of a prototype as a proof of concept for direct user input of the query, quality properties and priorities, covered in Chapter 6.

7. Demonstrate that the prototype performs appropriately according to the specified requirements and can provide qualitative information, which varies appropriately according to the context. See Chapter 7 for further detail.

1.6 Contributions to Research

We have addressed each of the above objectives during the course of this thesis. Through this, we show how the development of the Data Quality Manager is based on

research already carried out in an isolated mode [Naumann00], [Burgess03b], [Gertz04] but provides a novel approach to the utilisation of that previous approach.

The DQM is novel because:

- It proposes the management of representation and assessment of data quality when data is integrated from multiple sources with known provenance.

- The identification of a set of usable quality criteria in a generic Reference Model is novel. The Reference Model classifies data quality properties considering different user perspectives. Most of the existing classifications of data quality are context related or focused on specific user perspective.

- For the assessment of data quality at different levels of granularity and for derived data utilising the reference model, users specify the context of the query by selecting which criteria they wish the calculations to be based on.

- The facility to provide users with qualitative information relative to the originations of data, and the elements from which data was compound are also novel.

- To be able to bring together the different qualitative measures and produce a single score which can be applied at different levels of granularity utilising known approaches by user selection.

- This work has been done within heterogeneous multi-database environment.

1.7 Thesis Outline

The remainder of this thesis is organised as follows:

Chapter 2 Background

This chapter presents an extensive analysis of the present work concerned with data quality and resolution of data inconsistencies. To date, none of them have measured and assessed data quality at different levels of granularity, nor for derived data. An important feature was to develop expressive data quality models and tools that help users and Information Technology (IT) managers to capture and analyze the state of data quality within an Information System (IS).

Chapter 3 The Data Quality Manager

This chapter describes a conceptual framework for data integration considering the Data Quality Manager (DQM) followed by the discussion of the Reference Model, the Measurement Model, and the Assessment Model, which make up the DQM.

We present a Generic Quality Reference Model, based on existing research that summarizes data quality properties from different user perspectives, includes uniqueness as a quality criterion, and provides a general structured classification.

The Measurement Model contains identified metrics from existing research. Such metrics have been extended to measure at database, relation, and attribute levels of granularity for original data sources.

We identified a new assessment-oriented classification based on the level of granularity assessed: direct assessment and indirect assessment.

Chapter 4 Data Provenance

The Data Provenance Chapter illustrates the implementation of an algorithm for extracting data provenance and its associated quality properties from each ancestor to help the assessment processes as an extension of the Assessment Model.

Within the Data Quality Assessment Model, we present methods of assessment at different levels of granularity for a number of data quality properties such as accuracy, uniqueness, currency, timeliness, volatility and response time.

The Data Quality Assessment Model considers Data Provenance as a helpful mechanism for the assessment of data quality of derived data.

The Assessment Model adds assessment by provenance to the classification identified in Chapter 3.

Chapter 5 Multi Attribute Decision Making (MADM)

The MADM Chapter discusses the MADM problem in terms of data quality criteria, its scores and priorities. We next discuss the scaling criteria methods, the specification of

quality priorities by weighting normalization, a brief analysis and evaluation of the available ranking methods. The chapter concludes with some recommendations regarding which ranking method to use based on the scaling methods.

The outcome of this chapter is the provision of a mechanism to provide a quality score for data at different levels of granularity within a system, which is novel.

Chapter 6 Design and Implementation

The purpose of this chapter is to identify the requirements and to consider them for the design, and the implementation of a prototype as a proof of concept of our hypothesis.

Chapter 7 Test and Experimentation

This chapter tests the potential capabilities of the prototype by implementing the TPC-C and TPC-H benchmark database suites. The objective was to demonstrate that the prototype performs appropriately in terms of achievement of requirements, and can provide qualitative information, which varies appropriately acording to the specification of context.

Chapter 8 Conclusions and Future Work

The chapter reviews the outcomes of the previous chapters, explaining how this thesis has addressed the objectives along with the contributions to research, conclusions, DQM limitations, and future work.

Chapter 2 Background

2.1 Introduction

This thesis addresses two main areas: On the one hand, Data Integration is concerned with matching, merging or linking data from a variety of disparate sources. On the other hand, Data Quality is concerned with measuring, profiling, correcting, standardizing, verifying, and improving data.

The aim of this chapter is to establish the context and background regarding the problem of extensional inconsistencies during data integration as well as how data quality can help to deal with them. Therefore, the first section will introduce a summary of Data Integration on Heterogeneous Systems. We then proceed to the discussion of a number of data quality definitions, classifications, metrics, and assessment methods from previous research, according to their different perspectives, application domains and limitations. We next present how data quality has been addressed during data integration, and finally this chapter concludes with the lessons learned from previous work and the potential areas of development in terms of our research.

2.2 Data Integration in Heterogeneous Databases

Data integration is the process of extracting and merging data from multiple heterogeneous sources to be loaded into an integrated information resource [Batini86].

Database integration is divided by Motro and Rakov [Motro98] into two main problems, intensional and extensional inconsistencies. Intensional are related to resolving the structural and schematic differences between the component databases. Extensional inconsistencies are related to reconciling the data differences among the participating databases.

Solving structural, syntactical and semantic heterogeneities between source and target data has been a complex problem for data integration for a number of years [Batini86], [Sheth90], [MacKinnon98], [El-Khatib00].

One solution to these problems has been developed through the use of a single global database schema that represents the integrated information with mappings from global

schema to local schemas, where each query to the global schema is translated to queries to the local databases using these mappings [Batini86].

The use of domain ontology, metadata, transformation rules, user and system constraints have resolved the majority of the problems of domain mismatch associated with schematic integration and global schematic approaches. However, even when all the mappings, semantic and structure heterogeneities are solved in the global schema, extensional consistencies may not have been achieved, because the data provided by the sources may be mutually inconsistent. At the same time, each autonomous component database deals with its own properties or domain constraints on data, such as accuracy, reliability, availability, timeliness and cost of data access [Angeles04].

Several approaches to solve inconsistency between databases have been implemented:

1. By reconciliation of data, also known as data fusion: different values become just one using a fusion function (i.e. average, highest, and majority), depending on the data semantic [Motro98].

2. Based on individual data properties: associated with each data source (i.e. how recent, complete or correct it is). These properties can be specified at different levels: the global schema design level, the query itself or in the profile of the user [Anokhin01].

This second approach is the subject of this thesis and consequently, an extensive review of definitions, classifications of data quality criteria, along with their corresponding metrics and measurement methods are presented in the following sections.

2.3 Data Quality Definitions

The subjective nature of the term Data Quality (DQ) has allowed the existence of general definitions such as *"fitness for use"* in [Wang96], which implies that quality depends on customer requirements.

The definition established by Redman et al in [Redman96], suggests that data quality can be obtained by comparing two data sources *"A datum or collection of data X is of higher or (better) quality than a datum or collection of data Y if X meets customer needs better than Y"*.

Another definition is *"The distance between data views presented by an Information System and the same data in the Real World"* in [Wand96], which means that quality depends on the capacity of an information system to represent facts of the real world. Consequently, careful handling of data shall be done during its life cycle.

Recently, data quality has been defined as *"the capability of data to be used effectively economically and rapidly to inform and evaluate decisions"* [Karr05]. Such definition considers data quality not as the end but the means for making informed decisions.

However, these definitions are not very useful when data quality requires to be evaluated. Consequently, data quality rather than being defined has been characterized by multi attributes or dimensions according to specific application domains, types of assessment or customer requirements for instance, that shall be accomplished in order to be suitable for use.

As the determination of data quality is by comparing its corresponding attributes [Cavano78], [Redman96], this collection of attributes must be defined, classified, measured and compared in order to determine an overall quality. However, quality properties are often of a quantitative or qualitative nature, the former being easy to measure, but not the latter, which are subject to personal expertise. Furthermore, *"..What may be considered good quality information in one case (for a specific application or user) may not be sufficient in another case"* [Huang99], which means that even defining the quality attributes, and identifying their corresponding measures and assessment methods, the overall quality will depend on the specific priorities given by data consumers.

2.4 Data Quality as a Multidisciplinary area

The handling of Data quality is a multidisciplinary area, which has been addressed by a number of research disciplines such as management, statistics, and computer science The management researchers focus on control strategies for data manufacturing systems [Wang95], [Strong97]. The statistical research started with mathematical theories for considering duplicates in statistical datasets, then with projects such as NISS, which is related to the measurement error and survey methodologies, data editing and record linkage [Karr05]. The Computer Science perspective is concerned with how to *"define, measure and improve the quality of electronic data, stored in databases, data*

warehouses and legacy systems" [Scannapieco02] with projects such as DaQuincis detailed in [Scannapieco04], and FusionPlex explained in [Anhokin03].

2.4.1 The Management perspective

The Massachusetts Institute of Technology (MIT) and the Cambridge Research Group, among other institutions, have co-founded the MIT Total Data Quality Management program (TDQM) [TDQM]. The aim of TDQM is to create a theory of data quality based on disciplines such as Computer Science, Statistics, and the Total Quality Management field, and is focused on the definition and measurement of data quality, the identification and analysis of data quality impact, and the redesign of business practices and implementation of new technologies to improve information quality.

In Total Data Quality Management the concepts, principles and procedures are presented as a methodology, which defines the following continuous life cycle: define, measure, analyze and improve data as essential activities to ensure high quality, managing data as a product.

2.4.2 The Statistical perspective

The National Institute of Statistical Sciences (NISS) [NISS] and the Carnegie Mellon University among other institutions are currently working on the project "Data Confidentiality, Data Quality and Data Integration for Federal Databases: Foundations to Software Prototypes". NISS is proposing a cross-disciplinary research (Management, Computer Science and Statistics), for the development of theory, methodology, and software prototypes that will be applied to actual statistical databases, to deal with privacy, protection, confidentiality, and high-quality statistical data.

According to Alan F. Karr in [Karr05] statisticians have approached the consequences of poor data quality in different ways, mainly for data cleansing purposes such as:

- Statistical data editing, an automated process of stepping through the data records and correcting them if they violate pre-specified constraints.

- Measurement Error and Survey Methodologies: which are concerned with the modes of collection, interviewers, surveys, designs, and detection and evaluation of errors.

- Probabilistic record linkage, which is concerned with methods for evaluating the likelihood that pairs of records in different files are semantic matches.

2.4.3 The Computer Science perspective

The Database Management Systems have approached data quality at an isolated level mainly within relational databases at the management of transactions and the ACID properties, data entry level, database design, duplicate removal and record matching techniques. As Information Systems have become more complex, Data mining and Knowledge base systems are aimed to develop techniques that "discover" the quality "automatically" with only limited human guidance [Motro98]. Computer Science researchers consider the data quality problem as how to define, to measure and to improve the quality of electronic data, stored in databases, data warehouses and legacy systems [Scannapieco02].

From our point of view, such disciplines are complementary. First, the definition of relevant quality properties is required. Followed by the specification of strategies to control the manufacturing processes, such as design, representation, and instantiation of data. The next step could be the detection of possible errors, by the application of mathematical theories to develop measurement and assessment methods. Finally, the identification of mechanisms to improve data quality, in order to maintain a pre-specified level of data quality.

The subject of this thesis from the Computer Science perspective is concerned with the identification, measurement, and assessment of data quality of derived data, and data sources at database, relation, attribute levels of granularity within multi-database environment to provide ranking of data sources based on the user specified context.

2.5 Data Quality classifications

This section is concerned with the identification and discussion of the quality properties available from previous research.

2.5.1 An Ontologically based approach

Richard Wang et al. in [Wang95] first proposed a definition of quality dimensions, and a framework for the analysis of data quality as a research area, based on ISO9000 series. Yair Wand et al. one year later [Wand96] developed an ontologically based approach.

This model analyzed data quality based on discrepancies between the representational mapping from the real world (RW) to the Information System (IS) and vice versa, through design and operational activities involved in the construction of an information system as an internal view. From an external point of view, an information system is considered as a black box with the functionality required for representing the real world system specified by the user.

A real world system is properly represented if there exists an exhaustive mapping, and no two states in RW are mapped into the same state in IS. Four intrinsic data quality dimensions were identified: complete, unambiguous, meaningful and correct.

Additionally mapping problems and data deficiency repairs were suggested. The analysis produced a classification of data quality dimensions as related to the internal or external views. However, data quality measurement methods were not addressed (See Table 2.1).

	Dimensions
Internal view (design operation)	**Data- related:** Accuracy, reliability, timeliness, completeness, currency, consistency, precision **System-related:** Reliability
External view (use, value)	**Data-related:** Timeliness, relevance, content, importance, sufficiency, usableness, usefulness, clarity, conciseness, freedom of bias, informativeness, level of detail, quantitativeness, scope, interpretability, understandability . . **System-related:** Timeliness, flexibility, format, efficiency

TABLE 2.1 CLASSIFICATION BASED ON INTERNAL OR EXTERNAL VIEW FROM [WANG96]

2.5.2 Data Quality for the Information Age

Thomas Redman considers data as a triplet of conceptual model, data value, and data representation, and their correspondence with the data life cycle model activities ("define view", "obtain values" and "present results") [Redman96]. The set of data quality properties derived from each step of these activities, might help on the internal and external perspectives of data quality.

Therefore, such classification is suitable for use across different application domains at the data level. As we are not interested in any specific context, such classification could be considered for our generic Reference Model. See Table 2.2 for further detail. Nevertheless, this approach did not address measurement methods.

Conceptual View		
Content	Relevance	Obtainability
	Clarity of Definition	
Scope	Comprehensiveness	Essentialness
Level of Detail	Attribute Granularity	Precision of Domains
Composition	Naturalness	Identifiability
	Homogeneity	Minimum unnecessary redundancy
View Consistency	Semantic Consistency	Structural consistency
Reaction to Change	Robustness	Flexibility
Value		
	Accuracy	Completeness (entities and attributes)
	Consistency	Currency/Cycle Time
Representation		
Formats	Appropriateness	Format Precision
	Efficient use of Storage	Interpretability
	Format Flexibility	Portability
	Ability to Represent Full Values	
Physical Instances	Representational Consistency	

TABLE 2.2 CLASSIFICATION BASED ON CONCEPTUAL, VALUE, AND REPRESENTATION ASPECTS OF DATA [REDMAN96]

2.5.3 The Data-Consumer Perspective

A different classification of data quality dimension was developed by Diane Strong et al. in [Strong97]; it is based only for a data-consumer perspective and the data quality categories were identified as intrinsic, accessibility, contextual, and representational, where each category was directly addressed to different data quality (DQ) dimensions (See Table 2.3).

DQ Category	DQ concerns	Causes	DQ Dimensions
Intrinsic	Mismatches among sources of the same data are common cause of intrinsic DQ concerns	Multiple sources of same data. Judgment involved in data production.	Accuracy Objectivity Believability Reputation
Accessibility	Lack of computing resources. Problems on privacy and confidentiality Interpretability. Understandability. Data representation	Systems difficult to access. Must protect confidentiality. Representational DQ dimensions are causes of inaccessibility.	Accessibility Access Security
Contextual	Operational Data production problems: Changing data consumers needs. Distributed computing.	Incomplete data. Inconsistent representation. Inadequately defined or measured data. Data results not properly aggregated.	Relevancy Value Added Timeliness Completeness Amount of Data
Representational	Computerizing and data analyzing	Data inaccessible because: Multiple interpretations across multiple specialities and limited capacities to summarize across image.	Interpretability Ease of understanding Concise and Consistent representation Timeliness

TABLE 2.3 CLASSIFICATION BASED ON DATA-CONSUMER PERSPECTIVE, FROM [STRONG97]

However, the previous classification contains quality properties relative to the data consumer only, and it includes quality properties relative to the accesibility provided by the system.

2.5.4 Conceptual, Logical, and Physical Perspectives

The aim of Jarke et al. in [Jarke98] was to develop mathematical techniques for measuring data-warehouse quality aspects. He proposed three perspectives to manage data-warehouse quality and to classify data quality dimensions.

The conceptual perspective corresponds to a business model of the information systems of an enterprise.

The logical perspective focuses on the actual data models involved. Therefore, a model consists of a schema, and a schema is composed of types. The physical perspective is concerned with the physical components in a data-warehouse architecture. On the one hand, there are agents who act as programs that control other components or transport data from one location to another; on the other hand, there are data stores or databases. The Data usage quality classification is with regard the usage and querying of data, and it is shown in Table 2.4.

Data usage Quality	Logical Perspective		Physical Perspective	
Accessibility	Is the schema definition accessible by the users?	Is the type visible and accessible for others?	Is the network sufficient for delivered data?	Is the data store accessible?
Availability	Frequency of updates	Frequency of updates	Response time	Uptime of data store, response time
Security	Level of security (access rights)	Level of security (access rights)	Are there physical access restrictions?	Is the store able to prevent unauthorized access?
Usefulness	Is the schema used by any user?	Is the type used by any user?	Is the data delivered by the agent really used in the destination store?	Is the data in this store queried by a user?
Interpretability	Is the schema understandable	Is the type understandable?	Is the data delivered understandable?	Is the data stored understable?

TABLE 2.4 EXAMPLES FOR MEASUREMENT TYPES FOR DATA USAGE QUALITY DIMENSIONS [JARKE98]

The data quality classification refers directly to properties of the stored data (not schemata or models). Therefore, is related just to the physical perspective (Table 2.5.)

Data Quality	Physical Perspective	
	Agent	Data Store
Completeness	Number of tuples delivered wrt. expected number	Number of stored null values where there are not expected
Credibility	Believability in the process that delivers the values	Number of tuples with default values
Accuracy	Number of delivered accurate tuples	Level of preciseness; Number of accurate tuples
Consistency	Is the delivered data consistent with other data	Number of tuples violating constraints, number of coding differences
Data Interpretability	Number of tuples with interpretable data, documentation for key values, is the format understandable?	Number of tuples with interpretable data, documentation for key values, is the format understandable?

TABLE 2.5 EXAMPLES FOR MEASUREMENT TYPES FOR DATA QUALITY DIMENSIONS FROM [JARKE98]

2.5.5 The assessment oriented model

Information Quality (IQ) criteria have been classified in an assessment-oriented model by F. Naumann in [Naumann00], where for each criterion an assessment method is identified.

Individual users determine the scores of subjective criteria based on their experience, knowledge, and focus. Therefore, subject assessment is recommended in case of experienced users.

The scores of objective criteria are determined by a careful analysis of data. The process criteria are derived by querying process.

In this classification the user, the data, and the query process are considered as sources of information quality by themselves, (see Table 2.6).

Object-criteria and process-criteria shall be utilized for an unbiased assessment of data, and for any level of user experience. Therefore, this classification shall be considered for our assessment methods within the identfication of our Assessment Model.

Assessment Class Source IQ of metadata	IQ Criterion	Assessment Method
Subject-Criteria	Believability	User experience
	Concise representation	User Sampling
	Interpretability	User sampling
	Relevancy	Continuous assessment
User	Reputation	User experience
	Understandability	User sampling
	Value-added	Continuous assessment
Object-Criteria	Completeness	Continuous assessment
	Customer Support	Parsing, sampling
	Documentation	Parsing
	Objectivity	Expert input
Information/ Data	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert input
Process-Criteria	Accuracy	Sampling, cleansing
	Amount of Data	Continuous assessment
	Availability	Continuous assessment
Query Process	Consistent representation	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

TABLE 2.6 AN ASSESSMENT-ORIENTED CLASSIFICATION FROM [NAUMAN00]

2.5.6 Classification for Cooperative Information Systems

This classification contains the same quality properties proposed by Wang and Strong in [Wang96] plus some others proper to the Collaborative Information Systems (CIS) and it is based on the elements required for their measurement as Naumann in [Naumann00].

The subject dimension corresponds to all quality properties whose measurement requires the user expertise.

The object dimension considers all proper quality characteristics that require data per se to be evaluated.

The architectural is composed by dimensions related to the CIS architecture identified in [Cappiello02], see Table 2.7.

Category	DQ dimension	Elements required for measurement
subject level :	interpretability, ease of understanding, concise representation (brief and complete, compactly represented) and accessibility;	User expertise
object level	believability, accuracy, objectivity (unbiased) , reputation, representational consistency, internal consistency (schema related) and completeness (at attribute and value level)	Data
architectural level	availability, responsiveness, source availability, source responsiveness.	Architecture of Cooperative Information Systems
Process level	relevance, timeliness, appropriate amount of data, process completeness, value added, access security, history (annotations on transformation functions) and cost (how much the errors due to bad quality cost).	Context and cooperative processes

TABLE 2.7 THE DAQUINCIS ASSESSMENT-ORIENTED CLASSIFICATION FROM [CAPPIELLO02]

2.5.7 Product and Service Performance/Information Quality

The AIM Quality Methodology (AIMQ) [Yang02] is a practical methodology for assessing and benchmarking IQ organizations, with three elements. In the first place, there is a Product and Service Performance/Information Quality (PSP/IQ) Model, described in [Kahn02], which presents a quality dimension classification by product quality and service quality using the information consumer perspective, and consolidates the dimensions into four quadrants: sound, dependable, useful, and usable information. These quadrants are relevant to IQ improvement decisions. In the second place, an IQA instrument measures IQ for each dimension. In a pilot study, using questionnaires answered by information collectors, information consumers, and information systems professionals in six companies, these measures were averaged for the four quadrants and the scale used in assessing each item ranged from 0 "not at all" to 10 "completely". In the third place, the IQ Gap Analysis Techniques assess the organization information quality for each of the four quadrants against the IQA benchmark, to identify IQ issues, which are the bases for focusing IQ improvement efforts.

This methodology uses questionnaires as the main measurement method, taking a very pragmatic but subjective approach.

2.5.8 NISS Project: Process, Data, and User Dimensions of Quality

"To become a science DQ must have a foundation built on measurement" [Karr05].

The process dimensions are those data quality properties related to the generation, assembly, description and maintenance; the Data quality properties are those associated with data themselves; and the User dimension quality characteristics are those related to use and users. The User and Process dimensions are subject to qualitative assessment, whether the Data dimension contains quality properties suitable for quantitative assessment see Table 2.8.

Category	DQ criterion
Process	Metadata, reliability, security, confidentiality.
Data	Record level: Accuracy, completeness, consistency, validity
	Database level: Identifiability and Joinability
User	Accessibility, Integrability, Interpretability, Rectifiability, relevance, timeliness.

TABLE 2.8 DQ CLASSIFICATION ACCORDING WITH PROCESS, DATA AND USER

2.5.9 Summary

The differentiation between Information Quality (IQ) and Data Quality (DQ) is most of the time implied. Therefore, the terms data and information are often used synonymously [Strong97], [Wang98]. DQ however, is related to quality at the early stages of information, such as accuracy, consistency and integrity characteristics at schema, data representation, and data value level [Redman96], and IQ is concerned with data quality in context, and how the information is produced and interpreted, and processed in some manner [Strong97], [Wang98]. We consider that as data quality is related to the early stages of information, data quality should be considered before information quality. Consequently, in order to consider that information has a good reputation the analysis of quality shall measure if data is correct rather than reputable. Good reputation is not a warranty of completeness, or accuracy. On the contrary, data quality shall be the cause of good reputation not vice versa.

Data quality is characterized by quality criteria or dimensions such as accuracy, completeness, consistency, and timeliness in several approaches such as [Wand96], [Motro98], [Gertz98], [Naumann02], [Strong97], [Pipino02] mainly because classification facilitates the characterisation and definition of an overall quality.

However, according to [Wang95] and [Ballou98], there is no general agreement on data

quality dimensions. For instance in the case of completeness, it has been referenced at record, attribute and query levels and consistency might vary according to either the representation or data value perspective. For a more detailed description of discrepancies among data quality definitions refer to [Scannapieco02].

We can conclude that the differences among concepts and classifications rely mainly on the different focuses according to the role and experience of the user. Therefore, a data quality classification that comprises data representation, data value, context from the design, by the product and customer based perspectives, from an internal and external focus is required.

2.6 Data Quality Assessment and Measurement

This section is concerned with the identification and discussion of the current assessment and measurement methods according to our research interest.

2.6.1 Positive and Negative Criteria

"The more the better" is a positive direction for some quality criteria [Naumann02], [Burgess04]. For instance, more completeness derives higher quality. However, this is not the case with all criteria, such as cost, where the minimum value is the best. However, from the provider perspective the higher the price the better, but from the consumer perspective a lower price is the best option.

2.6.2 Interval Scales

In order to assess, to record and to manipulate the identified criteria, it is important to define the nature of the assessment.

Quantitative criteria are those whose metrics are described by numeric or absolute values [Naumann03]. This is the case for consistency or accuracy, which value can be obtained from the ratio between the number of consistent or correct tuples and the total number of tuples.

In contrast, qualitative criteria such as reliability or believability are not easy to represent. Their values are "high" or "low". They can be represented using scalar formats such as ratio, (where distances are set in respect to a pre-established value), ordinal, (allowing us to state quality parameters in rank order), and intervals (to provide

equal distances between values and indicate the corresponding distance between attributes) [Hwang81].

2.6.3 Temporality

"Values of quality criteria may vary over time" [Burgess04]

When data changes frequently, new information about its quality should be stored regularly in the metadata [Gertz98]. When a quality criterion can still be used for some time after the data was recorded, this quality is static such as completeness, consistency and accuracy.

In the case of dynamic criteria such as update frequency or currency, these values of quality may vary over time.

2.6.4 Assessment methods

2.6.4.1 Objective Assessment

Objective assessment may use metrics with no consideration of the context application, or may use task dependent metrics, which include the organization's business rules, regulations, and constraints provided by the database administrator, to be applied to any data set [Pipino02].

- **Cleansing techniques**

In order to correct, standardize and consequently, to improve data quality, data cleansing has emerged to define and determine error types, search and identify error instances, and correct the errors.

"Data cleansing is applied especially when several databases are merged. Records referring to the same entity are represented in different formats in different data sets or are represented erroneously. Thus, duplicated records will appear in the merged database. This problem is known as merge/purge problem." [Maletic00]

According to [Buchheit02] the most common methods utilised for error detection are:

a) Statistical methods through standard deviation, quartile ranges, regression analysis, etc. [Barnet84], [Bock98].

b) Clustering that is a data mining method to classify data in groups to identify discrepancies.

c) Pattern recognition based methods to identify records that do not fit into a certain specific pattern.

d) Association rules to find dependencies between values in a record [Maletic00].

Data cleansing is commonly performed in offline time, which is unacceptable for operational systems. Therefore, cleansing is often regarded as a pre-processing step for Knowledge Discovery in Databases and Data Mining systems during the Extraction Transformation and Load (ETL) process. However, it is still a very time consuming task, *"The process of data cleansing is computationally expensive on very large data sets and thus it was almost impossible to do with old technology"* [Maletic00]. A brief comparison between Data Cleansing approaches is detailed in [Muller03].

- **The Parsing technique:**

By considering the actual data or a metadata, it is possible to determine if a given string (in this case an entire tuple or an attribute) is an element of the language defined by the grammar. Accuracy is commonly assessed by this method.

The assessment of value consistency is calculated objectively by parsing or cleansing techniques.

- **Sampling:** Samples of data are considered appropriate for finding the score of the entire data source. This method is often used for completeness, and accuracy criteria.

- **Continuous assessment:** In case of dynamic criteria, quality assessment is executed at regular intervals. Continuous assessment is required for timeliness, response time, and availability criteria.

2.6.4.2 Subjective Assessment

Subjective assessment depends upon the user experience, the task at hand, and the use of questionnaires [Ballou98], [Wang98].

- **User experience:** Data quality is assessed depending on previous user experience and knowledge of the specific domain and data sources. For instance, reputation and believability are criteria suitable to be judged by user experience assessment.

- **User sampling:** A user will assess data by analysing several sample results. The user should be skilled enough to find appropriate and representative samples. In the case of interpretability of data, users find which attributes are more suitable for sampling than others are.

- **Continuous user assessment:** In the case where finding representative samples of data is not possible, the user needs to analyse every data, not only samples. That is the case of relevancy or amount of data.

- **Contract:** The assessment is performed depending on the terms of the contract of agreement between the provider and the data consumer, which is the case of price or cost of data [Naumann02].

One example of subjective assessment is the use of control matrices proposed by E. Pierce in [Pierce04], to audit the information products. The evaluation is in terms of how well they meet the consumer's needs, how well they produce information products, and how well they manage the life cycle of the data after it is produced. The information product manager shall perform the evaluation.

The columns of the control matrix utilised by E. Pierce are the list of data quality problems and the rows correspond to the quality checks or corrective process exercised during the information manufacturing process to prevent them. Each cell shall contain a rating that can have three different forms:

a) The values Yes or No, whether the quality check exists or not.

b) The category of effectiveness at error prevention, detection, or correction ranked as "low", "moderate", or "high".

c) A number to describe the overall level of assessment of the quality check's effectiveness.

Table 2.9 shows the direction, the nature of the assessment, their corresponding assessment classes, and the temporality.

DQ Criteria	+/-	Scale	Assessment Class	Assessment Method	Variation Over time
Ability to represent null values	+	Qualitative	Objective	User sampling Parsing, Expert input	Static
Accuracy	+	Quantitative	Objective	Sampling, cleansing Data Bashing	Static
Adaptability	+	Qualitative	Subjective	User assessment	Static
Appropriateness	+	Qualitative	Subjective	User assessment	Static
Availability	+	Quantitative	Objective	Cont. Assessment	Dynamic
Believability	+	Qualitative	Subjective	User experience	Static
Completeness	+	Quantitative	Objective	Cont. Assessment Counting methods	Static
Conciseness Rep.	+	Qualitative	Subjective	User sampling.	Static
Consistency	+	Quantitative	Objective	Cont. Assessment	Static
Consist. Representation	+	Qualitative	Objective	Parsing. Verification	Static
Currency	+	Quantitative	Objective	Cont. Assessment	Dynamic
Format Flexibility	+	Qualitative	Objective	Parsing, user samples	Static
Interpretability	+	Qualitative	Subjective	User sampling	Static
Portability	+	Qualitative	Objective	Migration Process	Static
Precision	+	Quantitative	Objective	Parsing, User sample	Static
Price	-+	Quantitative	Objective	Contract	Dynamic
Reputation	+	Qualitative	Subjective	User experience	Static
Response time	-	Quantitative	Objective	Cont. assessment	Dynamic
Timeliness	-	Quantitative	Objective	Parsing	Dynamic
Uniqueness	+	Quantitative	Objective	Cleansing tech.	Static
Usability	+	Qualitative	Subjective	User assessment	Static
Value Added	+	Qualitative	Subjective	Cont. assessment	Static
Verifiability/Testability	+	Qualitative	Objective	Expert input	Static

TABLE 2.9 DATA QUALITY ASSESSMENT SUMMARY

The Table 2.10 presents the different quality dimension definitions with the relevant factors on each dimension and the proposed metric by author.

Dimension	Concern	Factors	Metric
Accuracy	"Inaccuracy implies that the Information System (IS) represents a Real World (RW) state different from the one that should have been represented" [Wand96]	RW/IS states	a) Compare with real world. b) Data bashing: compare with other databases
	"Whether the data available are the true values (correctness, precision accuracy or validity)" [Motro98]	Data values	#Correct values / #total values
	"The extent to which data is correct and reliable" [Pipino02]		
Uniqueness	"Collection where an entity from the real world is represented once."		#duplicated tuples / #Total tuples
Coverage c(S)	Measure for the number of tuples a source stores. Probability that an entity of the world is represented in the source. [Naumann03]	Tuples	#entities represented / Universal world

Density d(s)	Density of an attribute: Measure for how well the attributes stored at a source are filled with actual (non-null) values. d(A)
	Density of the source: Average density over all attributes of the global schema d(S) [Naumann03]
		attributes	d(A)=non-null values / #total values
			d(S)=average(d(Ai))

Completeness C(s)	"Ability of an IS to represent every meaningful state of the represented real world system. Thus is not tied to data-related concepts such as attributes, variables, or values" [Wand96]
	"The extent to which data is not missing and does have sufficient breadth and depth for the task at hand" [Pipino02]
	"All values for a certain variable are recorded" [Ballou98]
	"The degree to which all data relevant to an application domain have been recorded in an information system." [Naumann03]
		RW/IS states
		Data model (table, row, attribute, classes)
		schema
		Column
		Population
		Coverage
		Density
			a) Process: Counting methods and checksums
			1 – (#incomplete items / #total items)
			C(S)=c(S)*d(S)

Timeliness	"The extent to which data is sufficiently up to date for the task at hand" [Pipino02] The degree to which the recorded data are up-to-date" [Gertz98]	Currency Volatility	*Max (0,1 - (# currency / #volatility))*
Currency	"How fast the IS state is updated after the real world system changes." [Wand96] Age: of data, when first received by the system. Delivery time: when data is delivered by the user. Input time: When data is received by the system [Pipino02] "Whether the data are up to date, reflecting the most recent values" [Motro98] Time interval between latest update and time it is used [Bovee01]	Age Delivery time Input time	Process: Query Age + delivery time – input time Time request – last update time
Volatility	"The rate of change of the real world" [Wand96] "Refers to the length of time data remains valid." [Pipino02].[Ballou98]	Time	Time data invalid, Time start valid Update frequency
Value Consistency	"Often referred as integrity constraints state the proper relationships among different data elements" [Pipino02] "The degree to which the data managed in an information system satisfy specified constraints and business rules." [Motro98].[Gertz98]	Values of data on Integrity constraints	Data edits 1 – (#inconsistent / #total consistency checks)
Representation Consistency	"The extent to which data is presented in the same format" as consistent representation [Redman96]	Physical rep. data	Verification of a data value format with a reference one. 1 – (#inconsistent repres. / #total consistency checks)
Believability	"The extent to which data is regarded as true and credible" [Pipino02]	S=Source of data A=Accepted stand. P=Prev. experience	Min(A,S,P)
Accessibility	"The extent to which data is available, or easily and quickly retrievable" [Pipino02]	TR=Time request TD=Time delivery TN=Time no longer useful.	*Max (0,* *1* – (TR – TD / TR – TN))

TABLE 2.10 QUALITY DIMENSIONS DEFINITIONS, DETERMINANT FACTORS AND METRICS BY AUTHOR [BALLOUU98],[BOVEE01],[GERTZ98], [GERTZ04], [NAUMANN03],[MOTRO98], [PIPINO02], [WAND96].

2.6.5 Summary

From the material presented in this section, we can conclude that in spite of the large number of data quality properties that have been identified, only a small number include a corresponding metric. The principal reason for this is that most of them are context and user dependent.

On the one hand, data quality is suitable for being assessed at representation, value, and context level. On the other hand, data quality can be of qualitative or quantitative scales, dynamic or static, positive or negative. Therefore, assessment of quality is a complex and frequently very domain specific task.

As data, users, and query processes are sources of quality information by themselves [Naumann00], data value quality criteria shall be more convenient to assess than data at representation and context level, because the prime is suitable for being assessed by analysis of data, or query process and the latter depend on the domain and experience of users.

According with the assessment-oriented classificaction identified by Naumann et.al in [Naumann00], the assessment of subjective data quality properties such as believability, reputation, and interpretability depends on the role of the user and their experience, making the task complex, individual, and biased. Therefore, as our intention is to address a full range of users, we can conclude that objective, user independent criteria evaluations will be more viable, and useful.

2.7 Measuring Data Quality in Heterogeneous Systems

The following section discusses the way previous approaches have dealt with data inconsistencies during data integration.

There are several important areas of related work to consider not only from research but also from industry perspectives.

2.7.1 Research approaches

2.7.1.1 Data Integration Techniques based on Data Quality aspects (1998)

Gertz developed some data integration techniques in [Gertz98] and [Gertz98b], based on data quality aspects within an object oriented data model, and data quality

information stored in a metadata. Quality aspects such as timeliness, accuracy and completeness were considered in the process of database integration. The main aspect was the assumption that the quality of the data stored at different sites can be different and the quality varies over time. Query language extensions were necessary to support the specification of data quality goals, such as "most accurate" or "most up to date" for global queries and thus data integration. In the case of data conflicts between semantically equivalent objects, the object with the best data quality must be chosen. If no conflicts exist between objects but their quality level is different, the integrated objects are grouped to indicate this situation to the global query interface. However, the quality goals specification limits the possibility of more combinations of priorities from the user, because they are not given in weights or percentages, just the "the most accurate" or "the most up to date". Consequently, not just one or two combinations of quality priorities will satisfy users. One result might be good enough for one user under an specific situation, but of poor quality for other.

2.7.1.2 Multiplex (1998)

The project MULTIPLEX directed by Motro and Rakov [Motro98], addressed the problem of extensional inconsistencies. MULTIPLEX was based on accuracy and completeness as quality criteria. This model assigned a quality specification for each instance of a relation, and these quality specifications were calculated by extending the relational algebra. The multi-database designer addresses the conflict resolution strategy, and the quality estimates. The quality of answers was calculated by the measure of arbitrary queries from the overall quality specification of the database. In the case of multiple sets of records as possible answers to one query, each set of records has an individual quality specification. A voting scheme, using probabilistic arguments, identifies the best set of records to provide a set of ranked tuples to the user, but no further information about their associated quality. Therefore, users are neither able to establish their quality preferences or priorities nor to take part in the resolution process.

2.7.1.3 Quality-driven Integration of Heterogeneous Information Systems (1998)

Information Quality reasoning is defined as the integration of information quality aspects to the process of selecting the best information sources for the optimization of query planning by considering user priorities [Naumann02]. The aim is to identify and rank high quality plans, which produce high quality results. However, such aspects are

related through the establishment of information quality criteria, assessment, and measurement methods, under the following assumptions. First, query processing is concerned with efficiently answering a user query to a single or multi database. In this context, efficiency means speed. Second, query planning is concerned with finding the best possible answer given some cost or time constraint. Query planning involves regarding many query execution plans across different, autonomous sources that together form the complete result. The query planning "finds plans that are correct, but possibly generate different results, while classical optimization considers plans that all produce the same result".

The plan selection process consists of three phases: In the first phase, information sources with poor quality are discarded before the execution of the query by using the Data Envelopment Analysis method (DEA) [Charnes78], according to "source-specific quality criteria" such as understandability, reputation, reliability and timeliness. In the second phase, the aim is to find all the correct query plans given by the combinations of query views of the global schema (named query correspondence assertions or QCA) [Naumann99] that provide correct answers to the user query. In the last phase, the best plans are determined by the "QCA specific quality criteria" [Naumann02] such as availability, price, accuracy, and response time and the "attribute specific quality criteria" such as the number of attributes and the completeness for each attribute for the ranking of the plans using the Simple Additive Weighting method (SAW) explained in [Hwang81].

The completeness of the query result derived from different sources is approached in [Naumann03] considering the number of results (coverage) and the number of attribute values in the result (density). Completeness is calculated as the product between the density and the coverage of the corresponding set of information sources. F. Naumann et al. consider subjective quality criteria, and suggest expanding personal profiles with their corresponding quality scores.

There is a classification of specific quality criteria according to the level of granularity (in this approach data sources, queries and attributes). However, there is no further specification of how to assess quality at different levels of granularity.

Data sources are ranked using the DEA method. Therefore, there is no consideration of user priorities for this process. Besides, subjective criteria are used for discarding data

sources such as reputation and understandability.

2.7.1.4 FusionPlex (2001)

An enhancement of the Multiplex system FUSIONPLEX [Anokhin01], [Anokhin03] stores information features or quality criteria scores in metadata. The considered quality dimensions are timestamp, accuracy, availability, clearance and cost of retrieval. Inconsistencies are resolved by data fusion, allowing the user to define data quality estimation as a vector of features weights, performance thresholds and a fusion function at attribute level, as required. This approach reconciles the conflicting values at attribute level using an intermediate result named polyinstance, which contains the inconsistencies. In the first place, the polyinstance is divided in polytuples, and using the feature weights and the threshold, members of each polytuple are discarded. In the second place, each polytuple is separated into mono-attribute polytuples using the primary key, assuming that the same value of the primary key between databases refers to the same object but with different data values, and attribute values are discarded based on corresponding feature values. Finally the mono-attribute tuples are joined back together resulting in single tuples. Extensions to the relational data model and relational algebra were required to execute global queries to establish user priorities over quality properties to resolve inconsistency. The main disadvantages of this approach are that different semantic versions of attributes can be fused in the solved tuple, and the data fusion process is costly.

2.7.1.5 Data Quality in Cooperative Information Systems (DaQuinCIS) 2002

The aim of DaQuinCIS was to define an integrated framework to improve data quality in cooperative environments [Mecella03], [Missier03].

Such a framework started from the TDQM methodology. Then it was extended to suit the cooperative information systems requirements with a methodology for data quality enhancement and a distributed architecture supporting data quality monitoring and improvement [Scannapieco04].

Quality properties such as interpretability, availability, history, cost, security, reliability, accuracy, completeness, consistency and timeliness were considered. The use of a metadata was required to store the quality score, the meaning of the quality value, and how the measurements were carried out.

As users were involved in a cooperative system, policies regarding quality requirements, security of information, quality monitoring and improvement were developed in [Marchetti03] and [Baldoni03].

An important element of the framework is the Quality Factory, which is divided into the quality analyzer, the quality assessment, the quality improvement, the quality monitoring and the quality certificate.

The quality analyzer is in charge of the specification of the required quality for the identified data.

The quality assessment involves the use of data, process and user.

The quality improvement verifies data according to the quality specifications and in the case of non-alignment provides a description of the problem found based on quality improvement patterns [Bertolazzi03].

The quality monitoring module stores the events in which data quality requirements are not satisfied in a file, to be used as a guide in the cleaning and improvement process. Consequently, the generation of a quality certificate was possible [Cappiello03].

The quality certificate contains two elements: the quality data with the score of each quality property, and security information. The latter is in turn, formed by the level of confidentiality of data being transferred, the authentication of the data source, which can be for trusted organizations or external organizations and a signature to guarantee the association between the data and its creator.

The main contribution of this approach is its contribution to the Cooperative Information Systems in terms of the ability to assess, control and improve data quality. This can be achieved because requirements, constraints, commitments, and data to be shared among the participating systems are commonly agreed. This approach takes into account the specification of data granularity as the combination of elementary data items that are subject to quality metrics. For example, an address is a compound of other attributes on the same record, and there is a difference between computing the quality of this aggregate and computing an aggregate indicator over a set of items. However, the rating methods and the meaning of the quality criteria depend on the data providers. In other words, the measurement is not only subjective but also different methods are

utilised to measure quality, yielding different results. Furthermore, data derived from multiple data sources is not considered, (due to the Presumption of Primary Authorship).

2.7.1.6 Using Multiple Quality Criteria to focus Information Search Results (2002)

This approach proposes a methodology to use quality criteria to help information searching, from the data consumer perspective and suitable to many application domains [Burgess02].

A "Generic Framework of Information Quality" was developed with around 60 information quality properties classified hierarchically according to time, utility and cost [Burgess03]. In order to prove the flexibility and extensibility of such a framework in any specific domain, two applications were developed: A Quality Toolkit to maintain the framework and an Information Search Environment (ISE) to use quality measures to focus on search [Burgess04].

The ISE considered Multi Attribute Decision Making methods for the ranking of the query results, such as the Simple Additive Weighting method (SAW) and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method. The latter was enhanced to use preference values for criteria called TOPSIS-GP [Burgess03b].

Three application domains were tested: UK Universities, Cars and Home Freezers. After the use of the ISE under the already specified domains, it was possible to prove that "changing the quality requirements for a search can result in statistically significantly different ranking order of the retrieved information items" [Burgess03b].

Nevertheless, this approach was focused on information search not on measurement and assessment of data quality involved in heterogeneous systems, taking the Presumption of Primary Authorship and the Presumption of Atomicity.

2.7.2 Industry approaches

2.7.2.1 Dataflux

Dataflux [Dataflux] is a data management company that proposes data profiling, data quality, integration, enrichment, and monitoring processes, and helps to redefine and validate business rules, detects data quality errors, and integration issues, but it is not

concerned with tools that can help with taking decisions. Unfortunately, these processes are long and expensive.

2.7.2.2 Evoke

Evoke provides the "ability to grade data on a row by row basis in order to assess its fitness for specific business processes" through "management-level reports that measure the six key attributes of data quality (completeness, conformity, consistency, accuracy, duplication and integrity) across all data". Evoke then, provides quality measures just at the row level [Evoke].

2.7.2.3 Trillium

Trillium [Trillum] can help to implement, integrate and build data quality into complex enterprise systems, optimize and fine-tune business rules, generate a data quality scorecard and extend by adding multiple countries or data from other business systems. "Trillium Software System makes it easy to customize business rules to meet unique organizational needs". Software System rules are stored in simple text files that can be replicated or ported across platforms to create Total Data Quality. The architecture proposed by Trillium depends on Heterogeneous Systems that are going to share their business rules and will be the same across the federation.

2.7.3 Summary

Several actions to measure, assess, and improve data quality have been taken until now such as data editing, correction and monitoring, not only by researchers but also by the industry such as Dataflux, Evoke, and Trillium through commercial data quality software tools. However, these processes are commonly long, expensive and limited [Maletic00].

Within these reviewed approaches, and others that have been researched, there is no consideration of derived data. Data in all of these considered as a product of a primary source. However, due to the explosion of information over the last decade, we cannot assume that any data source is necessarily the point of origin of the data users require. Hence, the fundamental presumption of current data management practice, the "Presumption of Primary Authorship" must be challenged.

Users should be provided with information regarding data as an atomic value, or if it is composed data, what the atomic values were and the quality generated from. This challenges the "Presumption of Atomicity".

The key for a company to success is by meeting customer expectations through high quality products. To improve quality, a company requires improving processes of definition, production, measurement, control, and support of their products, which is a demanding and costly task. It is worth mentioning that the underlying reasons for poor quality in organizations and the methods for rectifying this problem are not the subject of this thesis.

2.8 Conclusions

- Data Quality is related to the early stages of information, such as accuracy, consistency and integrity characteristics at schema, data representation, and data value level. In contrast, IQ is concerned with data quality in context, and how the information is produced and interpreted.

- From our perspective data quality (DQ) is not an end; it is the way to improve decisions based on data. Therefore, we propose to measure the relative quality of data to speed informed business decisions as a parallel alternative.

- The main objectives of this thesis are concerned with the identification of useful data quality properties for the measurement, and assessment of data quality of derived data, and data sources at multiple levels of granularity within heterogeneous multi-data source environment, to provide data consumers with qualitative information.

- As we want to support a full range from experienced to naive users, the assessment methods utilised should provide meaninful and useful scores. Therefore, objective criteria, and process criteria should be included in the Assessment Model which are user independent, rather than subjective criteria, which can determined by individual users based on their experience and background.

- A data quality classification that comprises data representation, data value, and data in context, from the design, product, and customer based perspectives, and

considering internal and external focuses is required for a generic and expresive model [Gertz98].

- The processes to measure, to assess, and to improve data quality have been addressed not only by researchers but also by industry. However, these processes are commonly long, expensive and limited [Maletic00].

- Regarding data inconsistencies during data integration, consideration of quality properties from the user perspective has not been fully taken into account.

- The fusion of data with different quality properties with no consideration of user perspective might result in more semantic problems.

- From the existing approaches, data has always been considered as correct and reliable on the basis that is the product of a primary data source. Therefore, no consideration of derived data has been approached until now. The qualitative information provided to the user should contain measures of quality, the original data sources where data comes from, and the components of integrated data.

- There is no consideration of the process of integration (i.e. data fusion, data replication, or data transformation) during data quality measurement and assessment. In other words, measuring quality of derived data has not been addressed.

- Very few approaches have considered quality properties at different levels of granularity on databases [Scannapieco04b], [Naumann04]. Not to mention levels of granularity within derived data.

Chapter 3 The Data Quality Manager

3.1 Introduction

In the previous chapter, we have shown that although there has been considerable past work in the resolution of data inconsistencies due to poor data quality over a number of years, expressive data quality models and tools remain to be developed [Gertz04]. Accordingly, this chapter describes the architecture of the Data Quality Manager, a conceptual framework for data integration considering the DQM, followed by the discussion of the first three elements of the Data Quality Manager.

3.2 Architecture

The Data Quality Manager (DQM) is composed by the following elements:

- Reference Model: It will contain the data quality criteria required for the measurement and assessment of data sources. The Data Quality Reference Model represents objective 1 of the present work and it is detailed in section 3.4.

- Measurement Model: It will contain the definition of the metrics required to measure data quality. The Data Quality Measurement Model represents objective 2 of section 1.5 and it is detailed in section 3.5.

- Assessment Model: The identification of methods of assessment is essential for a correct interpretation of the data quality indicators. This is the stated objective 3 of section 1.5. This model will be explained in sections 3.6, and 4.7 for the assessment of primary data sources, and the assessment of derived data respectively. The obtained scores will be stored in a Quality Metadata.

- Quality Metadata: A repository to contain the information required to map from the global schema to the local schema in order to resolve intensional inconsistencies within the multidatabase environment, it will also contain the quality scores per each data source. The Quality Metadata is described in section 6.4.4.

- Provenance Metadata: A repository to contain ancestors information for the tracking of provenance, which will be designed in section 4.4.

- Data Provenance: The process of tracking provenance of data sources which will be covered in section 4.5, and represents the objective 4 of this thesis.

- Ranking of Data Sources: The identification of Multi-Attribute Decision Making methods utilised for ranking data sources as the objective 5 of this work (detailed in Chapter 5).

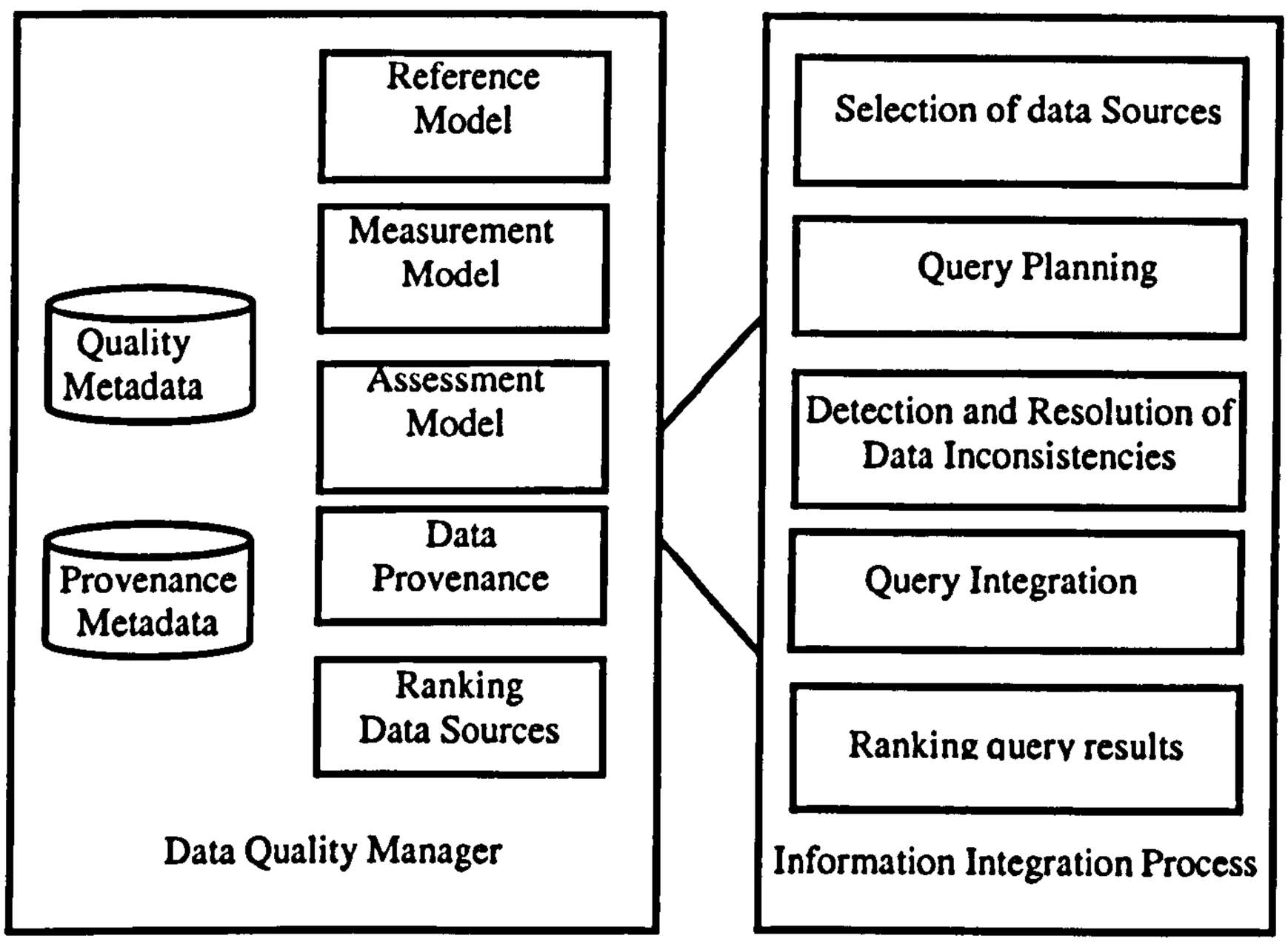


FIGURE 3.1 DQM ARCHITECTURE AND ITS RELATION WITH THE INFORMATION INTEGRATION PROCESS

The DQM could help the information integration processes (see Figure 3.1) such as the selection of data sources, query planning, data inconsistencies detection and resolution, query integration and the ranking of query results. The DQM is an element of the conceptual framework for information integration proposed in [Angeles04b] and it is explained in the following section.

3.3 Framework

The elements of the DQM could fit within the process of Data Integration, and its possible applications to deal with data inconsistencies are as follows:

1. The assessment of data sources has been divided into two steps:

a) The first step corresponds to the estimation of the quality scores of primary data sources, which will be stored in the Quality Metadata. See Figure 3.2.

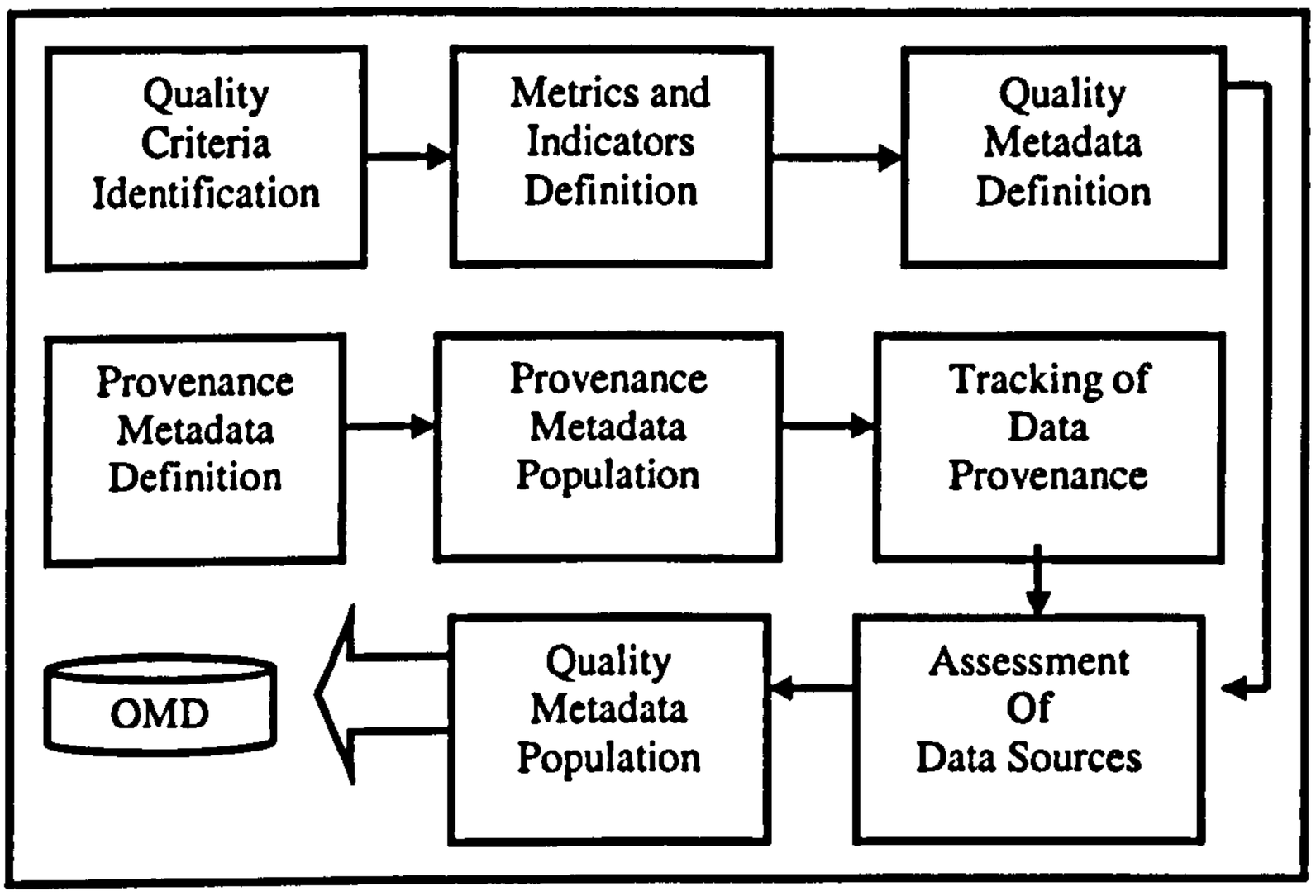


FIGURE 3.2 DQM: ASSESSMENT OF DATA SOURCES

b) The second step is the assessment of derived data, which requires the definition and population of a provenance metadata. The assessment is based on the quality scores of their corresponding ancestors. Figure 3.3 shows this process.

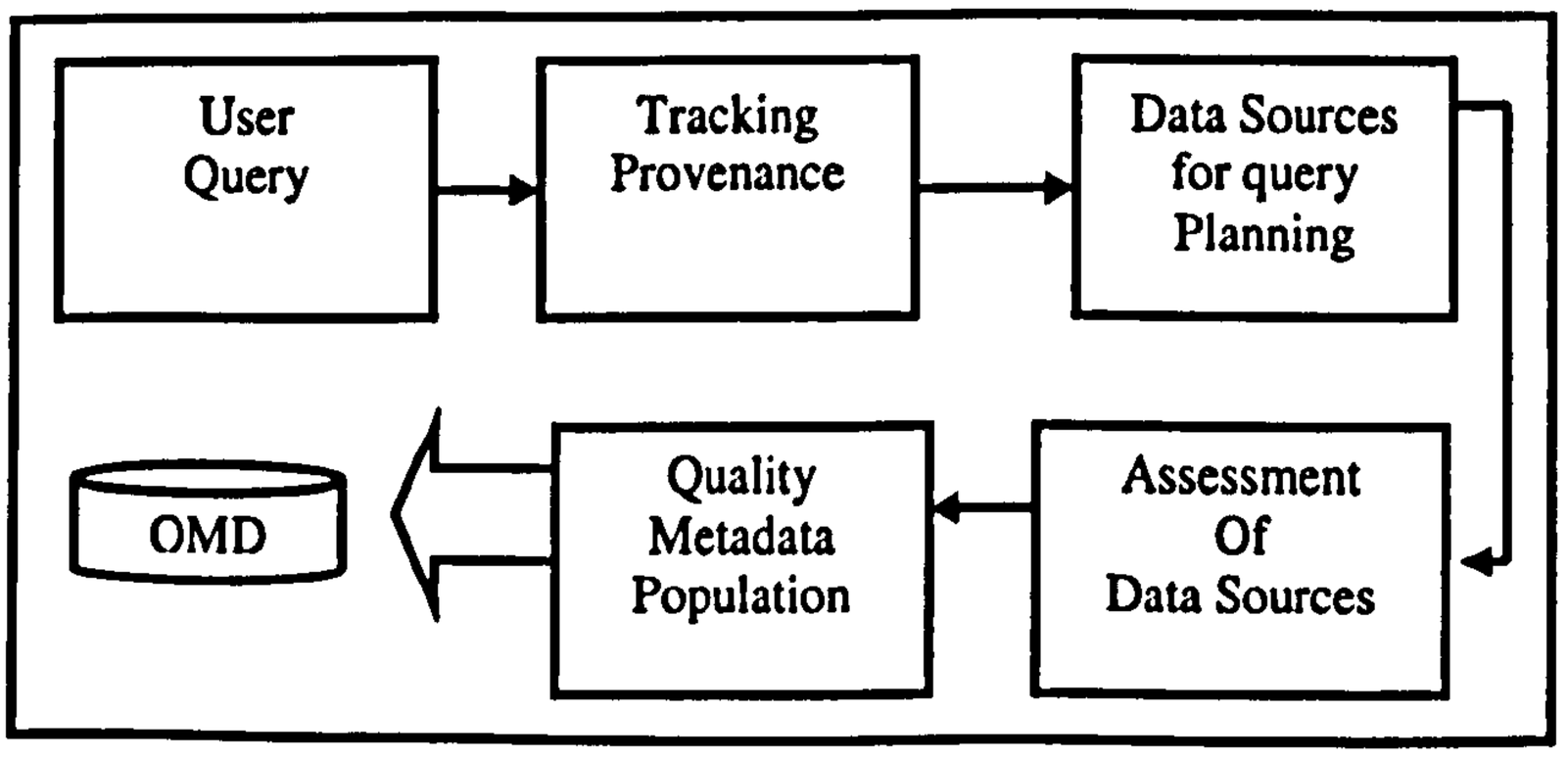


FIGURE 3.3 DQM: ASSESSMENT OF DERIVED DATA

2. The selection of the best data sources before the execution of the queries is on the bases of its quality scores (see Figure 3.4).

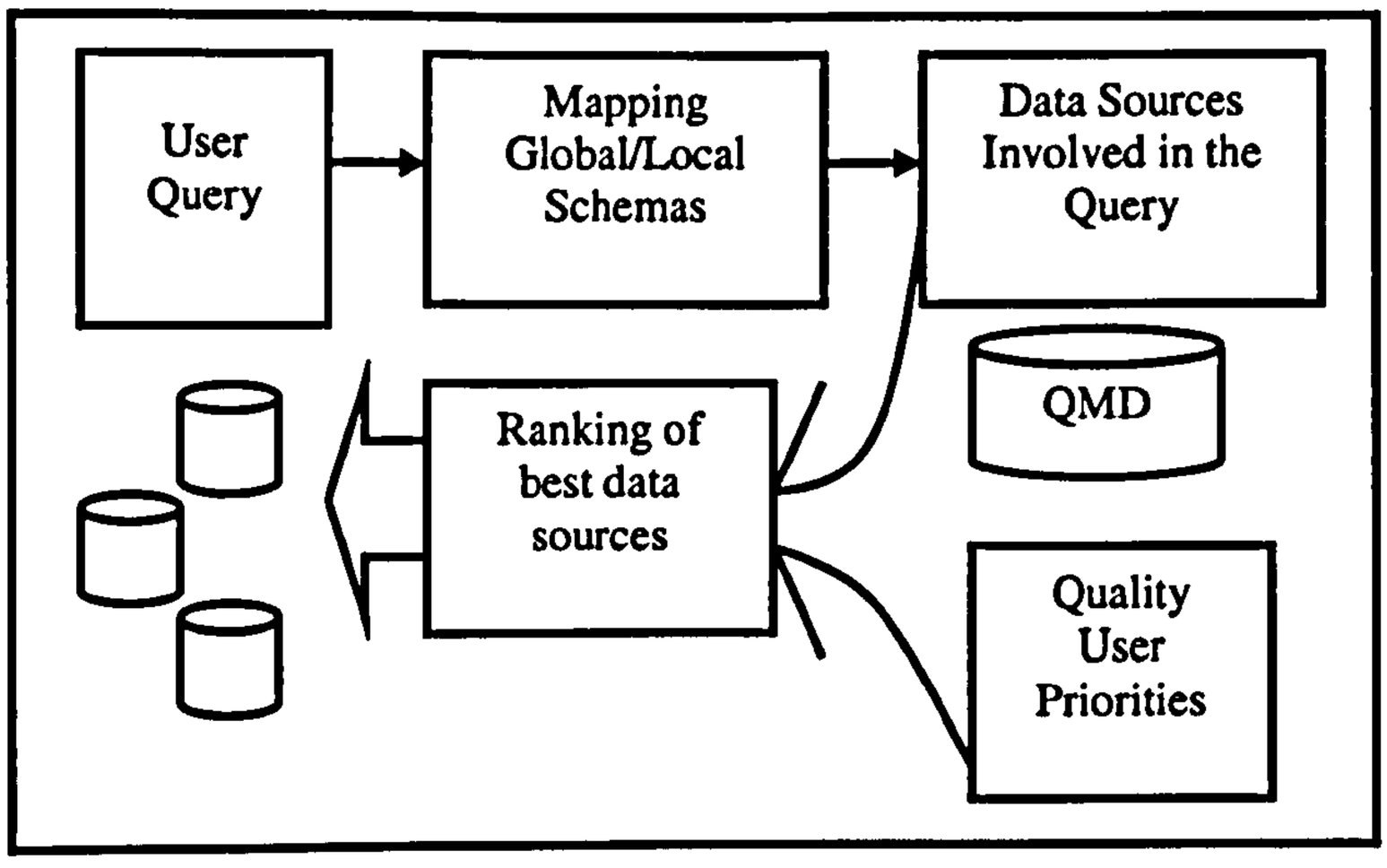


FIGURE 3.4 SELECTION OF BEST DATA SOURCES

3. The consideration of data quality scores helps the query planning by finding the best combination of data sources for the execution plan (see Figure 3.5).

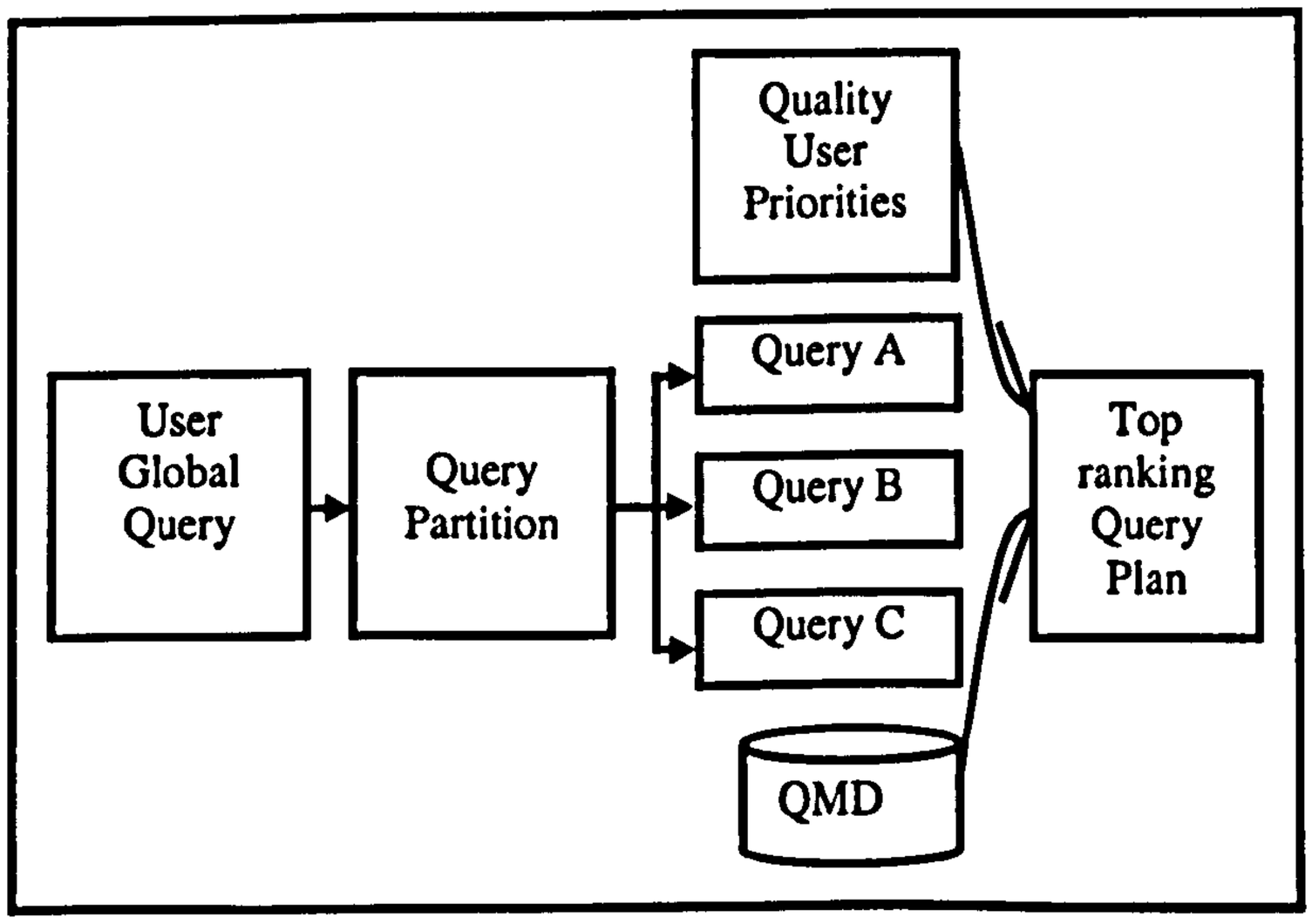


FIGURE 3.5 QUERY PLANNING

4. After query execution, and the detection of inconsistent data, data quality can be used to perform data fusion for the resolution of inconsistencies. However, care must be taken to avoid different versions of elements of fusion (see Figure 3.6).

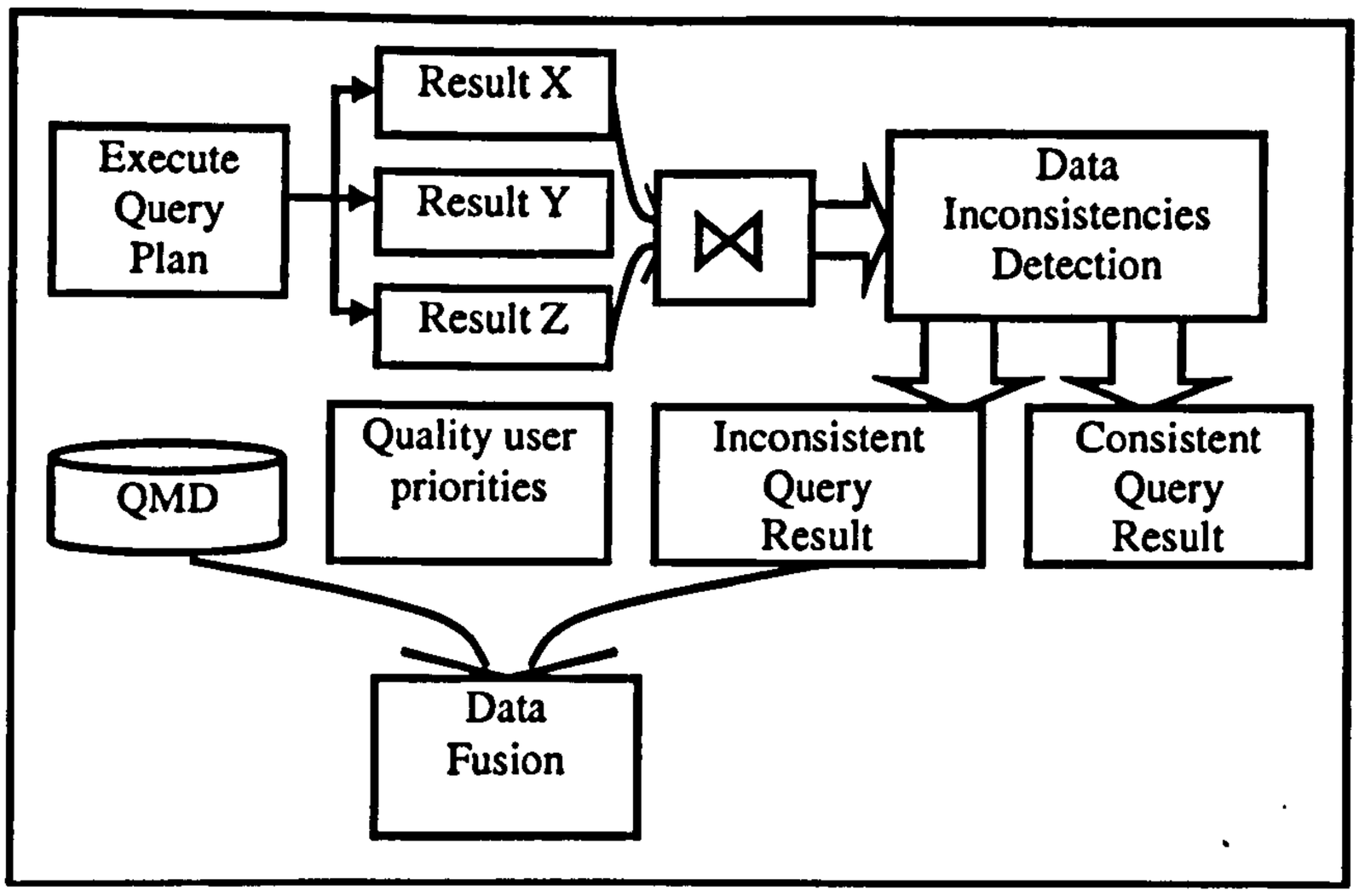


FIGURE 3.6 DETECTION AND RESOLUTION OF DATA INCONSISTENCIES BY DATA FUSION

5. After data fusion, the query result is integrated and presented to the user. Optionally the information presented to the user can be ranked considering the remaining query plans (see Figure 3.7).

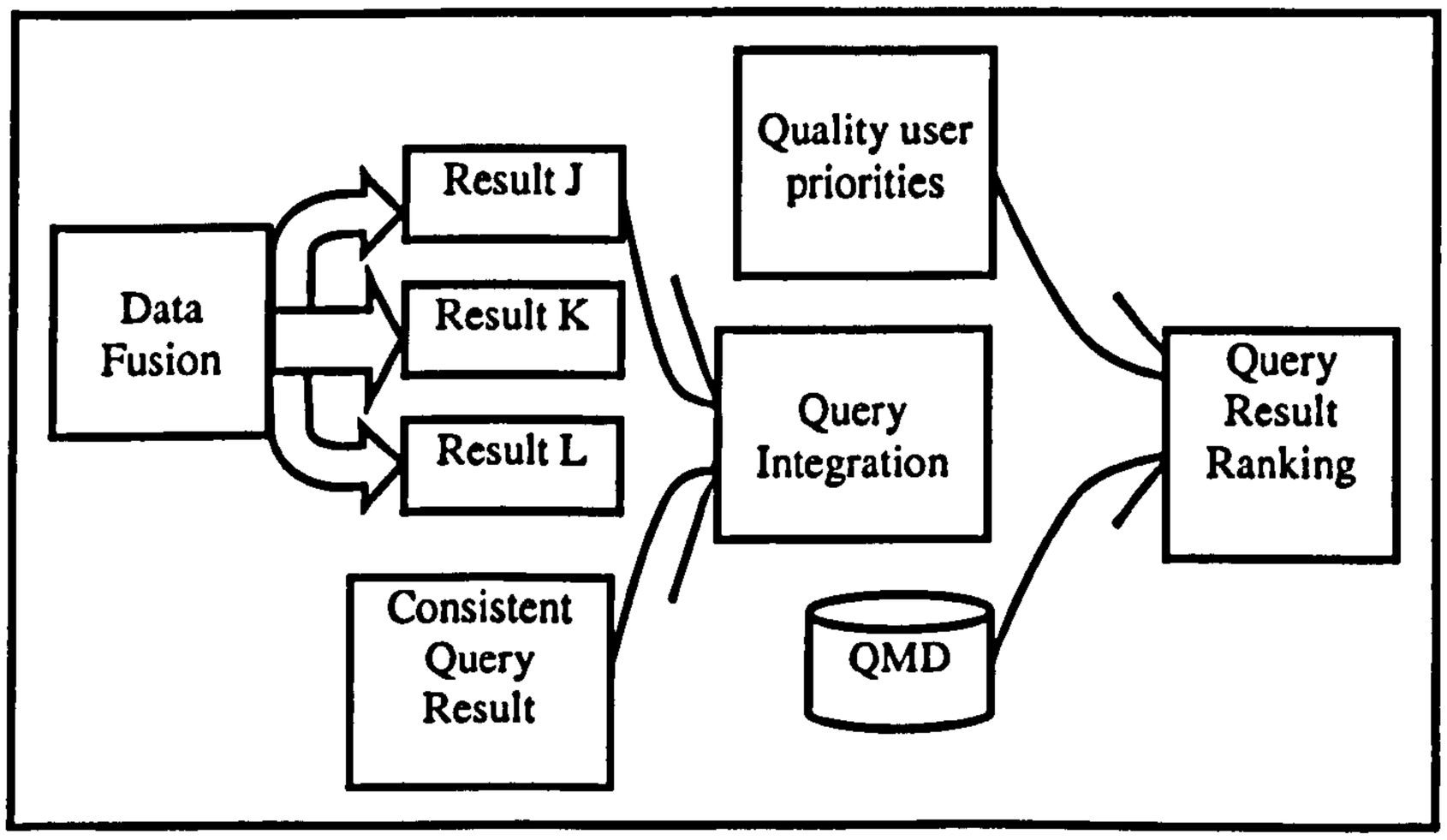


FIGURE 3.7 RANKING OF QUERY RESULT

6. In order to rank the data sources, the data quality scores previously stored in the metadata are used as a whole with their corresponding priorities. See Figure 3.8.

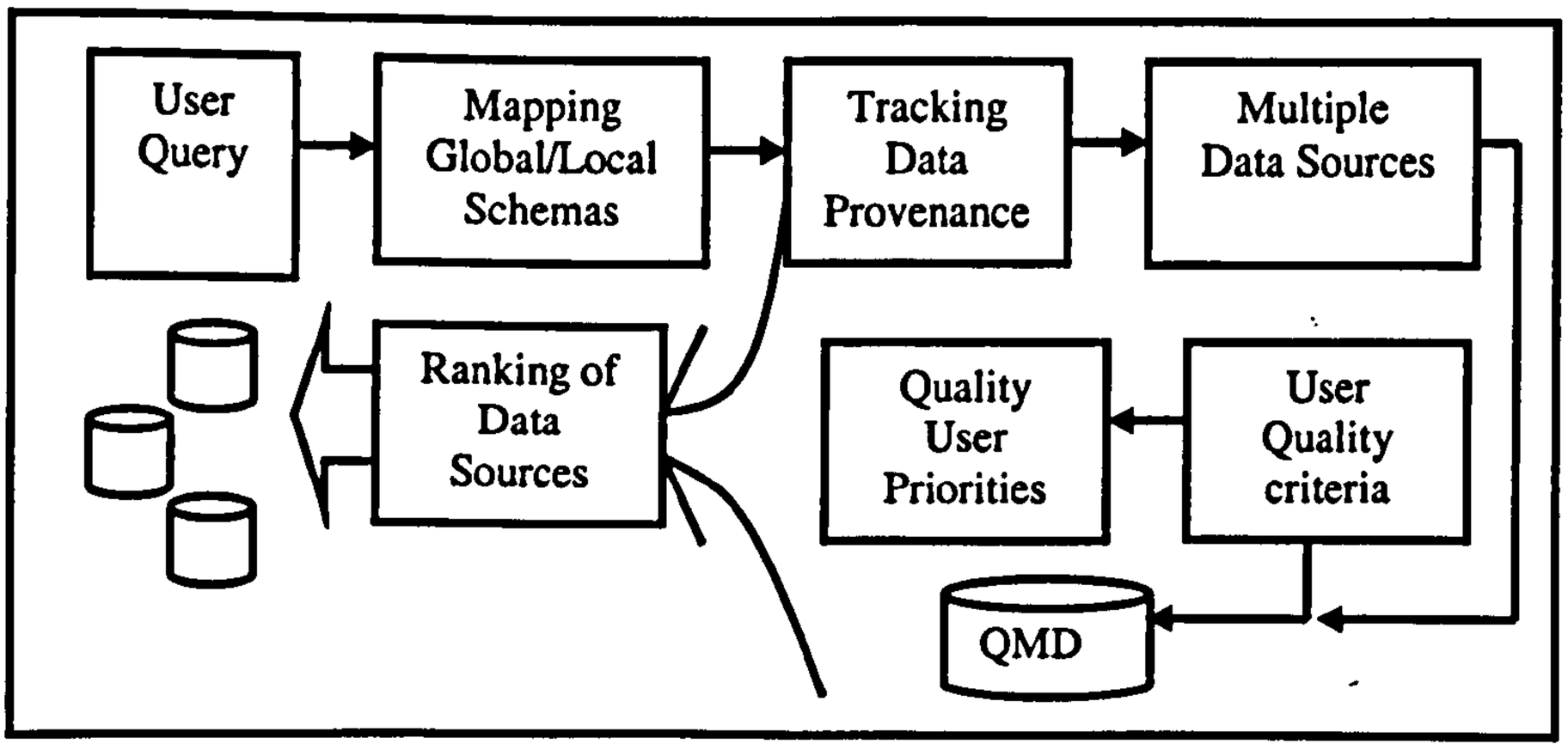


FIGURE 3.8 DQM: RANKING OF DATA SOURCES

The DQM could also be part of a diagnostic pre-process for data cleansing, or after data cleansing to evaluate data quality improvement.

The DQM represents a component of the conceptual framework. The DQM is designed to utilise data quality measures to provide qualitative information. As we have explained, such information could be further used within the data integration processes. However, this work concentrates solely on the design and development of the DQM. Having explained the framework, we will detail each component of the DQM through the rest of this chapter.

3.4 Reference Model

The step required to attain objective 1 is to conduct research into previous attempts to identify data quality dimensions from previous quality classifications from different perspectives and try to summarize them in a generic Data Quality Reference Model.

3.4.1 Classification

As we have mentioned in Section 2.3, there are many definitions of data quality under different focuses such as the data life cycle in [Redman96], or *"data as a raw material and information is the output"* in [Wang96]. Analysing data quality from the customer perspective, most web data consumers are aware of lack of accuracy and incompleteness of data. However, databases have traditionally been considered as sources of information that are precise and complete. Quality is relative to the customer expectations. Moreover, each user has different requirements depending on their profile, context, and knowledge. We require identifying a model that comprises different

focuses on data quality for any kind of users. The existing quality classifications are based on the data consumers focus on quality [Wand96], assessment-based [Naumann00], or context related [Jarke98]. This problem has already been detected by Gertz et.al. "..First concentrate on general and expressive models that can be used for modelling Data Quality metrics" [Gertz04]. Therefore, the Reference Model shall relate to different perspectives: The internal view is that data design activities are concerned with data value and data representation. It also should take into account the product-based perspective, and the external view from the customer perspective related to data context. This classification tries to undertake all the aforementioned perspectives.

The first step for the identification of the data quality properties was based on the literature review covered in Chapter 2 related to data quality properties. We have considered for each quality property the number of times it has been included within various approaches shown as "frequency" column in Table 3.1 which presents the resultant quality matrix.

	Redman	Wang 96	Strong 97	Motro 98	Jarke 98	Nauman 02	Pipino 02	Gertz 98	Karr 05	Frequency
Accuracy	•	•	•	•	•	•	•	•	•	9
Completeness	•	•	•	•	•	•	•	•	•	9
Interpretability	•	•	•		•	•			•	6
Accessibility		•	•		•	•	•		•	6
Consistency	•	•			•		•	•	•	6
Relevance	•	•	•			•			•	5
Timeliness		•	•			•	•	•	•	6
Amount of Data		•	•			•		•		4
F. Conciseness	•	•	•			•				4
F. Consistency	•	•	•			•				4
Reliability		•				•		•	•	4
Availability					•	•		•		3
Believability		•	•			•	•			4
Reputation		•	•			•				3
Understanding		•	•			•				3
Value Added		•	•			•				3
Currency	•	•					•			3
Usefulness		•			•					2
Rep. null values	•									1
Appropriateness	•									1
Cost						•				1
Use of storage	•									1
F. flexibility	•									1
F. Precision	•									1
Portability	•									1
Price						•				1
Response Time						•				1
Unbiased		•								1
Usability		•								1
Verifiability						•				1
Volatility		•					•			2
Uniqueness										0

TABLE 3.1 CONSIDERATION OF DATA QUALITY PROPERTIES BY AUTHOR.

There were projects with an internal and external focuses such as the ontological based approach project described in [Wang95]. There were also projects considering data as a triplet of conceptual, data value and data representation to describe data life cycle activities [Redman96]. Wang et al. for instance, identified a set of quality properties considering a data consumer perspective in [Wang96].

The second step was to identify each quality property in terms of the activity where it belongs. The aim was to re-organise the data quality properties in a general and flexible classification. For instance, the data quality properties relative to the data representation are originated at the design activity in an internal focus [Smart02]. Furthermore, the quality properties relative to the data value level perspective have been identified in product–based approaches, and the customer based perspective is concerned with data quality properties in context [Redman96].

Table 3.2 presents the Data Quality Reference Model, suitable to any application domain and supporting the full range from the internal focus to the external focus.

The Data Quality Reference Model contains all the data quality properties shown in Table 3.1. Besides, we have incorporated the quality property "uniqueness" to the model, because duplicated rows decrease quality in an isolated database.

The integration of data sources that contain duplicated tuples could result in extensional inconsistencies. Therefore, uniqueness should be included as a relevant quality criterion for the assessment of data quality to help in the resolution of extensional inconsistencies.

Existing approaches have not included uniqueness within data quality classification. In the case of relational databases, this situation may be allowed under the consideration that a relation is unique by definition, in spite of being a very well known issue in the implementation of Database Management Systems (DBMS) [Codd90].

As we have mentioned before, it is not our intension to analyse possible causes of flaws in data quality but to assess data quality.

Internal Focus	Design Level	Data Representation	Consistency Conciseness	
			Format	Appropriateness
				Understanding
				Format Precision
				Efficient use of storage
				Portability
				Interpretability
				Format flexibility
				Ability to represent null values
	Product Based	Data Value	Accuracy	
			Completeness	
			Consistency	
			Currency	
			Timeliness	
			Uniqueness	
			Volatility	
External Focus	Customer Based	Data Context	Reputation	
			Value Added	
			Response Time	
			Availability	
			Verifiability	
			Cost	
			Price	
			Believability	
			Usability	
			Usefulness	
			Reliability	
			Relevance	
			Unbiased	
			Accessibility	
			Amount of Data	

TABLE 3.2 DATA QUALITY REFERENCE MODEL FROM [ANGELES05B]

3.4.2 Concepts

As there is no agreement on the meaning of the data quality dimensions among different proposals because of its subjective nature and different focuses, see [Scannapieco02] for a detailed comparison, we present the data quality properties at the data value level with their corresponding concepts, see Appendix A for further detail on the rest of data quality concepts.

- *Accuracy* is the measure of the degree of agreement between a data value or collection of data values and a source agreed to be correct. [Lee04].

- *Completeness* is the extent to which data is not missing [Redman96], [Pipino02] and is divided by two quality dimensions: coverage and density in [Naumann03].

- *Consistency* is the extent to which the values are the same for overlapping entities and attributes. Data are consistent with respect to a set of constraints if they satisfy all constraints in the set [Redman96], [Motro98], and [Jarke98].

- *Currency* is the time interval between the latest update of a data value and the time it is used [Wang93], [Motro98].

- *Timeliness* is the extent to which the age of data is appropriate for the task at hand [Wang93], [Ballou98], and is computed in terms of currency and volatility. Timeliness has also been presented as context related dimension.

- *Uniqueness* is the extent to where an entity from the real world is represented once.

- *Volatility* is the interval of time where data remains valid on the system and is related to the update frequency [Ballou98], [Wang93].

3.5 Measurement Model

The Measurement Model is the stated objective 2 of this thesis. The aim of this section is to discuss which existing metrics are suitable for an unbiased, and user independent estimation of data quality scores to provide a more objective quality metadata [Gertz04].

As we are addressing any level of experience user, we require identifying a set of metrics that do not depend on the users to be evaluated (see section 2.8). Therefore, to establish which quality properties we should choose for unbiased, user independent measurement, the aim is to identify which measures correspond to an objective assessment according to Naumann's classification (referred to in Chapter 2, tables 2.9 and 2.10). Consequently, we have restricted our generic reference model in Table 3.2 to the instance perspective, because data quality properties at data value level are suitable for direct and objective measurement (see Table 2.10), and are the result of considering the product-based focus at data value level.

Most metrics proposed until now are just at one level of granularity. Particularly, completeness has been deeply approached in [Naumann03] and [Scannapieco04] with the coverage and density concepts in the former, and at different levels of granularity in the latter. However, we have taken into account not only attribute, and relation levels of

granularity following the completeness example given in [Scannapieco04b] but also the database level. We are considering the cardinality of a relation when measuring its quality. Therefore, the estimation of quality at database level is taken from the average score of its relations as a representative aggregation function. Discussion of aggregation functions is covered in Section 3.5.9 and Chapter 4.

As we have seen from the literature review, a small amount of quality properties includes a specific metric, because most of them are user and context dependent. The development of new metrics is not relevant for this research but to extend existing metrics to assess data quality at different levels of granularity (see objective 2 in section 1.5). Therefore, we have extended the existing metrics to assess data quality at database, relation, attribute, and tuple levels of granularity. In the following sub-sections, we mention the existing metric, and their corresponding reference for further detail. After that, we explain how we have extended the metric at value level, attribute level, tuple, relation, and database levels of granularity.

The strictness is a weak or strong characterization depending on evaluating the quality property as a percentage or as a Boolean function respectively [Scannapieco04b]. The strong characterization of the quality metrics is useful in applications in which it is not possible to admit errors at the corresponding level of granularity. For instance, in the case of accuracy at tuple level, it would be useful only and only if all the instances of its attributes are accurate.

In the case of the assessment provided by the DQM we have considered the weak strictness to make possible the comparison of data sources for a number of data quality properties. However, there might be alternatives where strictness could depend on the level of quality required, according to specific applications.

The metrics mentioned in this section consider the possibility of two assumptions in the relational model, namely the Closed World Assumption (CWA) [Reiter78]. The CWA assumes that the tuples in a relation are all and only the tuples that satisfy the relational schema, so its knowledge of the world is complete. The Open World Assumption (OWA) assumes that the tuples in a relation are a subset of the tuples satisfying the relational schema, so its knowledge of the world is incomplete.

Within the OWA is required the concept of the reference relation. The reference relation is the relation that contains all the tuples representing the real world, also called Universal Relation in [Naumann00]. Let R be the cardinality of the reference relation. However, reference relations are rarely available, the cardinality of the relation S namely n, is easier to obtain and utilised for practical purposes. Therefore, the CWA, allows to obtain an estimation of the quality measures in terms of what actually exist in each data source.

While the value and tuple metrics are valid in both assumptions, attribute and relation metrics have different definitions on the two models. Therefore, the definition for attribute and relation, both in the weak and strong strictness are proposed within the CWA. However, such definitions can easily adapted to the OWA. If R is equal to n, they coincide with the CWA case. If R is greater than n, then the definitions are the same as the ones provided in the CWA case but replacing n with R [Scannapieco04b]. Taking the OWA or the CWA would depend on the possible estimation of the cardinality of the reference relations.

Notation: Consider a set *Dm* of domains where each attribute a_1 drawing its values from d_1, a_2 from d_2,... a_m from d_m. A relation *S* is a finite subset of the Cartesian product of one or more domains $d_1 X\, d_2 X ... X d_m$ with *m* attributes. The relation *S* is abbreviated as $S(a_1{:}d_1, a_2{:}d_2 ... a_m{:}d_m)$. Each element of S has the form $d_1,d_2,\ldots,d_m$ and is called tuple t of relation S. The degree *m* of the relation *S* is the number of attributes in it. The cardinality *n* of the relation *S* is the number of tuples in it. The number of relations in a database D is denoted by *w*.

3.5.1 Accuracy

The accuracy metric is defined in [Motro98], and [Lee04] as the ratio between the correct values and the total values in the data source. Refer to Table 2.9 for the corresponding metric.

- Accuracy at value level $A_t^{a_i}$ corresponds to the presence of the correct data value within an specific attribute a_i in a tuple t , and is set by the following notation:

$$A_t^{a_i} = 1 \quad if \quad value \quad in \quad a_i \quad is \quad correct$$
$$A_t^{a_i} = 0 \quad otherwise$$

- Weak tuple accuracy $A_w(t)$ is the number of correct instances of attributes in a tuple t divided by the degree of the relation.

$$A_w(t) = \sum_{i=1}^{m} \frac{A_i^{a_i}}{m}$$

- The accuracy of a tuple t is strong $A_s(t)$ if all attribute values in a tuple t are correct.

$$A_s(t) = 1 \quad if \quad A_i^{a_i} = 1 \quad \forall i \in [1..m]$$
$$A_s(t) = 0 \quad otherwise$$

- Weak accuracy at attribute level $A_w(a_i)$ is the number of tuples with correct values for a specific attribute a_i divided by the cardinality of the relation S.

$$A_w(a_i) = \sum_{j=1}^{n} \frac{A_{i_j}^{a_i}}{n}$$

- The accuracy of an attribute i $A_s(a_i)$ is strong if all instances t_j of the attribute a_i in the relation S are correct.

$$A_s(a_i) = 1 \quad if \quad A_{i_j}^{a_i} = 1 \quad \forall j \in [1..n]$$
$$A_s(a_i) = 0 \quad otherwise$$

- Weak relation accuracy $A_w(S)$ is the number of tuples where every attribute is correct divided by the total number of rows.

$$A_w(S) = \sum_{j=1}^{n} \frac{A_s(t_j)}{n}$$

- Strong relation accuracy $A_s(S)$ is that when all the tuples contain correct values in every attribute, or when a relation contains strong tuple accuracy, and strong attribute accuracy.

$$A_s(S) = 1 \quad if \quad A_s(t_j) = 1 \quad , \quad \forall j \in [1..n]$$
$$A_s(S) = 0 \quad otherwise$$

- Then accuracy at database level $A(D)$ can be derived from the average of all accuracy scores at relation level.

$$A(D) = \frac{\sum_{k=1}^{w} A(S_k)}{w}$$

We are considering the cardinality of a relation when measuring its quality. The estimation of quality at database level is taken from the average of the scores of its relations as a representative aggregation function. Discussion of aggregation functions is covered in section 3.5.9 and Chapter 4.

3.5.2 Completeness

Regarding completeness, we have taken the corresponding metrics of [Scannapieco04b] and [Naumann03] for the value, attribute, and relation granularity levels, and we have incorporated completeness at the database level.

- *Coverage:* This is the measure for the number of tuples a source stores; in other words as the probability that an entity of the world is represented in the source [Naumann03]. This is also contemplated under the Open World Assumption without nulls completeness case, at the relation level of granularity, refer to [Scannapieco04b] for further detail. Coverage of a source c(S) will be defined as follows:

$$c(S) = \frac{n}{R}$$

- *Density of an attribute*: $d(a_i)$ is the measure of how well the attributes stored at a source are filled with actual (non-null) values (columns), in [Naumann03], a weak attribute completeness $C_w(a_i)$ case under the Closed World Assumption with nulls in [Scannapieco04b].

$$d(a_i) = C_w(a_i) = \sum_{j=1}^{n} \frac{C_{i\,j}^{a_i}}{n}$$

- *Density of the source d(S):* according with [Naumann03], is obtained by the average density over all density attributes.

- Weak relation completeness $C_w(S)$ is the number of tuples with all its attributes filled with non-null values divided by the number of tuples [Naumann03].

$$C_w(S) = \sum_{j=1}^{n} \frac{C_w(t_j)}{n} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} \frac{C_{t_j}^{a_i}}{m}}{n}$$

- The completeness at database level C(D) will correspond to the average completeness of its corresponding relations.

$$C(D) = \frac{\sum_{k=1}^{w} C(S_k)}{w}$$

3.5.3 Consistency

Data value consistency is the extent to which the values for overlapping entities and attributes are the same. Data is consistent with respect to a set of constraints if they satisfy all constraints in the set (refer to Table 2.10 for further detail). In the case of a composite primary key, the score is computed by the average of the multiple instances of its attributes.

- A value in an attribute a_i in a tuple t is consistent if and only if it obeys the corresponding constraints.

$$Cn_t^{a_i} = 1 \quad if \quad the \quad value \quad in \quad a_i \quad is \quad consistent$$
$$Cn_t^{a_i} = 0 \quad otherwise$$

- The weak consistency at the tuple level $Cn_w(t)$ is the number of instances of the attributes that are consistent divided by the degree of the relation.

$$Cn_w(t) = \frac{\sum_{i=1}^{m} Cn_t^{a_i}}{m}$$

- A tuple has strong consistency $Cn_s(t)$ if and only if all attribute instances are consistent.

$$Cn_s(t) = 1 \quad if \quad Cn_t^{a_i} = 1 \quad \forall i \in [1..m]$$
$$Cn_s(t) = 0 \quad otherwise$$

- The weak consistency at the attribute level $Cn_w(a_i)$ is the number of tuples where the instance of an attribute a_i is consistent divided by the cardinality of the relation.

$$Cn_w(a_i) = \frac{\sum_{j=1}^{n} Cn_{i_j}^{a_i}}{n}$$

- An attribute in a relation is strong consistent $Cn_s(a_i)$ if and only if all the instances of attributes *i* in the relation are consistent.

$$Cn_s(a_i) = 1 \quad if \quad Cn_{i_j}^{a_i} = 1 \quad \forall j \in [1..n]$$
$$Cn_s(a_i) = 0 \quad otherwise$$

- The weak consistency at the relation level $Cn_w(S)$ is the percentage of tuples with all instances of the attributes consistent.

$$Cn_w(S) = \frac{\sum_{j=1}^{n} Cn_s(t_j)}{n}$$

- A relation has strong consistency $Cn_s(S)$ if and only if all its tuples contain just consistent instances of attributes.

$$Cn_s(S) = 1 \quad if \quad Cn(t_j) = 1 \quad \forall j \in [1..n]$$
$$Cn_s(S) = 0 \quad otherwise$$

- Weak consistency at data base level is the average of all the consistencies at relation level across the database.

$$Cn(D) = \frac{\sum_{k=1}^{w} Cn(S_k)}{w}$$

The following time related metrics (covered in See Table 2.10), are considered at tuple level as the direct level of granularity where the measure could be obtained. The level of granularity of the estimation will depend on the DBMS.

3.5.4 Currency

The currency is the time interval between latest update and time it is used [Bovee01], [Wang93].

$$Cu(t) = \text{Time Request - last update time}$$

3.5.5 Response Time

Is the delay between the user request and the reception of the complete response from the Information System.

$$RT(t) = \text{Time Reception - Time Request}$$

3.5.6 Volatility

Volatility is the interval of time where data remains valid on the system, and it is related to the update frequency [Ballou98]. This dimension characterizes data according with the Information System. For instance, Data Warehouse systems have a very low volatility or no volatility at all, and Transactional systems contain very volatile data.

$$Vo(t) = \text{Update frequency}$$

3.5.7 Timeliness

This measure involves not only the currency of data but also if data is in time for an specific usage, and is given under the following terms.

$$T(t) = \max\left(0, 1 - \frac{Cu(t)}{Vo(t)}\right)$$

3.5.8 Uniqueness

A relation is unique by definition. However, Relational Database Management Systems (RDBMS) require user intervention to cope with uniqueness.

As uniqueness concerns with the number of tuples represented just once in a relation, this score is applicable at tuple, relation and database levels and the corresponding metrics are as follows:

- A tuple is unique if it is represented once in the relation.

  $U(t_j)=1$ *if tuple j is represented once in a relation*
  $U(t_j)=0$ *otherwise*

- The weak uniqueness at relation level $U_w(S)$ is the number of non-duplicated tuples divided by the cardinality of the relation.

$$U_w(S)=\frac{\sum_{j=1}^{n}U(t_j)}{n}$$

- Considering strong strictness for uniqueness, we can say that a relation S is strongly unique $U_s(S)$ just in case there are no duplicated tuples in it.

$$U_s(S)=1 \quad if \quad U(t_j)=1 \quad \forall j \in [1..n]$$
$$U_s(S)=0 \quad otherwise$$

- We can estimate uniqueness at database level by the average of its corresponding relation scores.

$$U(D)=\frac{\sum_{k=1}^{w}U(S_k)}{w}$$

3.5.9 Summary

In order to assess data quality at different levels of granularity, we have extended the metrics identified from previous research (covered in Chapter 2), and utilised the measures provided at lower levels of granularity to determine aggregated scores as we move through the levels of granularity as it is shown in [Scannapieco04b].

For instance, accuracy is first measured at the instance level of the attribute, then at the attribute level, then at the tuple level, then at the relation level and finally at the database level. The measurements are given by the aggregation of values at each of these levels as they are moving on. As a measurement of data quality is directly related

to the level of granularity, we conclude that scores measured at lower level of granularity will provide a greater degree of accuracy than aggregated scores produced at higher levels.

The functions utilised for aggregation of scores are commonly average, maximun, and minimum [Naumann00]. The appropriateness of an aggregation function will depend on the optimistic, conservative, or pesimistic approach taken according with the application context. It is not our intension to identify the best aggregation function, because there is not an absolute value. As long as the aggregation function reflects the user needs and it is consistently used, it should be enough for the estimation of quality and comparison purposes.

3.6 Assessment Model

This section is concerned with the identification of the processes required to represent, to interpret, and to assess data quality indicators, see objective 3 in section 1.5.

3.6.1 Temporality

Temporality relates to the frequency of change of data (referred to in Section 2.6.3). Therefore, in the case of static quality dimensions such as accuracy, completeness, consistency, and uniqueness, their corresponding scores shall be recalculated each time the data changes. In the case of dynamic criteria such as timeliness or currency, these values of quality may vary over time.

Temporality might depend on the application domain. On the one hand, if the source contains non-volatile data, which is the case of Data warehouses, the quality scores will not change, because data do not change. New scores corresponding to the new data added to the warehouse will be required and computed within a specific period, according to the Extraction-Transformation-Load (ETL) processes. On the other hand, in the case of very volatile data (such as Operational environments), the quality information requires refreshing after any update to data. The refreshing of quality information results in a heavy workload, making the task unviable. Therefore, the state of a tolerance period (time between the last updated data and the time the quality scores where computed) should be considered according to the specified requirements for practical purposes.

3.6.2 Traditional methods of Assessment

As this is a data-level approach, each quality property is assessed through query processes such as parsing, sampling, or continuous assessment as indicated in [Naumann00], and are represented in a quantitative manner. Some examples are explained as follows.

Accuracy is a positive quality property, whose assessment is by comparison.

Completeness assessment is by parsing data sources, through the coverage and density properties.

Currency, *response time*, and *volatility* time stamps depend on a number of hardware and software factors such as disk, transactions, lock level, and checkpoint/commit intervals the database manager applies. The lowest granularity the database manager applies during insert, update, and delete transactions are commonly given at record level. The highest granularity is at database level when these timestamps are stored in the log for recovery purposes. As these are directly time-related properties, they are quantitative amounts and interpreted as negative quality indicators.

As our purpose is to establish the relative quality among data sources, an exact measure is not necessary.

We can distinguish two types of assessment according to the level of granularity.

- *Direct assessment:* The process of assessment relates directly to the level of granularity such as uniqueness, which relates at the tuple level.

- *Indirect assessment:* The score is calculated based on other scores at other levels of granularity of the same source, such as accuracy at the relation level which value depends on accuracy at the row level.

3.7 Summary

From the literature review we conclude that despite a considerable research approaches on data quality, most of the existing classifications are context related or focused on a specific perspective. Therefore, we required identifying a general data-quality reference model that can be used for identifying a proper set of useful and meaningful data quality metrics. Gertz et.al has also detected this gap in [Gertz04].

We have detailed the three first elements of the Data Quality Manager.

- The Reference Model identifies data quality properties frequently considered as important characteristics of data. The DQRM is flexible enough for future extension, because it contains the different perspectives taken from previous researches. The Reference Model detailed in this chapter accomplishes objective 1 of this thesis.

- The Measurement Model takes its basis from existing and already used metrics for relational databases; it extends such metrics at database, relation, tuple, and attribute levels of granularity (objective 2 of section 1.5). From the metrics specification, we concluded that scores measured at lower level of granularity would provide a greater degree of accuracy than those measured at higher levels of granularity. Besides, the Measurement Model is suitable for future extension to include metrics under different data architectures.

- The Assessment Model is aimed at the evaluation of data sources within the perspective of a multi-database environment at different levels of granularity. Such a model is able to assess quality properties at database, relation, tuple, and attribute levels of granularity. As far as we know an assessment model with the above characteristics has not been approached before. The scores will be stored in the Quality Metadata for comparison purposes.

However, the assessment of integrated data is not possible using the current methods. Challenging the Presumptions of Primary Authorship, and the Presumption of Atomicity, a number of questions arise regarding assessment of quality: Where did the data come from? How did it get there? How old is it? What information do we need in order to analyse data quality? How can we compare between two data sources in order to decide which data to use? This led us to the consideration of the generation of derived data for its quality assessment, which is not a straightforward process.

Trying to answer these questions Chapter 4 will discuss the process of obtaining the ancestors of a data source namely Data Provenance (DP) in order to assess derived data. Assessment of data quality by considering data provenance represents an extension to our Assessment Model.

Chapter 4 Data Provenance

4.1 Introduction

Our consideration of Data Provenance arises from our interest in being able to set an assigned quality measure to derived data within a data source. In order to do this we need to know where the data has come from and how it has been changed on its passage from its original primary source to the data source we are investigating. Tracking of data provenance has already been developed by [Buneman01], [Tan04], [Bhagwat04].

With data provenace, we can determine the process by which derived data was produced by following the ancestor trail. Understanding this process provides information to assess derived data as stated in objective 4 of this thesis.

Our interest is in knowing where data comes from and how it has been changed during its life cycle. It considers the sharing of information across different applications, each of which can update a specific piece of data. Thus, after a number of changes, nobody actually knows who changed what data and under which conditions. Data changes all the time and under different conditions. Consequently, data becomes "dirty". A piece of data can be useful for a specific group and completely inaccurate for another group. Besides, data can become inconsistent considering that multiple sources with different data quality are available [Gertz04], [Motro98].

Therefore, in the case of derived data, the DQM should consider the tracking of data provenance along with associated data quality properties for a better approximation of quality.

The aim of this chapter is to show that extracting data provenance and its associated quality properties from each ancestor helps assessment processes.

We first discuss what data provenance is, we then identify the required annotations for tracking data provenance, and the structure of the metadata. After that, we explain the recursive algorithm developed to trace provenance. We next show how the quality properties associated with each ancestor can help to measure the quality of the data

derived. Finally, the last section concludes how data provenance helps users in the process of assessment of data sources.

4.2 Data Provenance concepts

"Data provenance is the description of the origins of a piece of data, and the process by which it arrived in a particular database." [Buneman01]

The process of tracking data provenance distinguishes between two perspectives named *where provenance* and *why provenance* by Peter Buneman et al in [Buneman01].

4.2.1 The where provenance

Where provenance is related to *"what pieces of input data helped create various values appearing in the output"* in [Buneman01] taking a syntactic approach. The process of *where provenance* aims to obtain the locations in the sources from which the data was extracted. On the one hand, the where provenance is the query itself if it contains hard-code values. On the other hand, it can be associated with one or more variables in the output expression of a query. In the case of derived data for example, *where provenance* may be presented as a transformation function involving elements of other data sources.

4.2.2 The why provenance

The analysis of *why provenance* allows understanding of the source data that had some influence on the existence of the data. *Why provenance* for relational databases has been approached by Woodruff in [Woodruff97] and Y. Cui et al. in [Cui00].

Why provenance is concerned with exploring *"what pieces of input data validate the existence of an output value, for a given query"* in [Buneman01].

According to W.C Tang in [Tan04], there are two approaches to compute data provenance.

4.2.3 Lazy approach

The lazy approach computes provenance of data only when needed, under the assumption that the transformation process is available and is given as a query. An example of lazy approach is given in [Buneman01].

4.2.4 Eager approach

The eager approach is performed by carrying the provenance of data along as data is transformed. An example of this approach is given in [Bhagwat04].

4.3 Problem description

Nowadays, information systems can provide a uniform and transparent user interface, enabling users to store and retrieve data in multiple data sources with a single query, even if the constituent databases are heterogeneous.

In this context, the number of information resources that can be used to solve one problem or to answer a query, can be huge. Furthermore, during the process of data integration data provenance is lost. Consequently, data consumers are not prepared to make correct decisions; they are overwhelmed with multiple sources that contradict each other due to poor quality.

When data providers do not create the data directly, but only collect or aggregate data, they are unable to determine the provenance of data; but they can annotate the ancestor, the one from which they are obtaining the data.

4.4 Annotations and the design of metadata

The next section will explain the use of annotations and the additional information required to achieve provenance.

4.4.1 Annotations

In this approach, annotations focus on the process by which data was included in a source, its ancestor, and the scores of its quality properties. Such annotations are stored in a metadata shared across the community, and maintained to support the data provenance process.

On the one hand, authors, data maintainers, and collectors are responsible for carrying out annotations such as the description of how data enters the database and its immediate ancestor. On the other hand, the Data Quality Manager estimates and stores the corresponding quality scores as explained in Chapter 3.

We next, identify some illustrative scenarios regarding the conditions under which suppliers or collectors introduce data into an Information System.

4.4.1.1 Data supplier creates data, according with the required representational consistency and domain, and has the authority to store it.

4.4.1.2 Data creator does not have the authority to store the data in the repository, but he asks the data administrator to store it for him.

4.4.1.3 Data supplier creates data, but the latter needs to be transformed or converted by the data administrator to resolve heterogeneity before it can be stored in the database.

4.4.1.4 Data supplier does not create data and it is not under a transformation process.

4.4.1.5 Data supplier does not create data and he transforms the data before storing it.

4.4.1.6 Data supplier does not create data, and data comes from several repositories with transformations on the way.

4.4.1.7 Data supplier does not create data, and data comes from several data sources by replication processes.

In the case of the scenario 4.4.1.1, the producer of data is directly supplying the data, and the relationship between the ancestor and data is just one to one. However, in the remaining scenarios, data is copied, transformed, or fused across a number of data sources; to suit data representation or semantics of different applications. These transformation functions are common for resolving intensional inconsistencies between heterogeneous systems. Therefore, the relationship between the ancestor and the data is many to one.

For instance, users are aware that replication is a copy data process only. Therefore, replication is a "less dangerous process" than transformation, which implies more manipulation of data, so we made a distinction by establishing 4.4.1.6 and 4.4.1.7.

Each scenario illustrates a possibility of decreasing data quality. Besides the more steps taken to represent data from the real world to an Information System, the more causes it would have to become dirty.

4.4.2 Design of metadata

The metadata required for provenance is designed under a relational data model. The reason for selecting a relational data model is because the author is familiarised with the design and implementation of relational data models.

From the above scenarios, we could identify the necessary entities. Therefore, the metadata consists of the entities *DataSourceInfo*, *Ancestors*, and *How_description*.

The entity *DataSourceInfo* contains information regarding the data sources involved in the federation, such as its unique identifier, physical location, name, capture process, its ancestor, and the function of transformation required to be generated, in case there is one. The entity *Ancestors* contains the identifiers of the data source of interest and the corresponding ancestors. The entity *How_description* contains the description of the process of data capture, already described in the scenarios.

In summary, when the cardinality of the relation between *DataSourceInfo* and Ancestors is just one to one, the *where provenance* is obtained directly from the *DataSourceInfo* entity. However, in case the cardinality of such relation is more than one, it is necessary to obtain all the ancestors from the *Ancestors* entity.

4.5 The process of tracking provenance

The algorithm of data provenance described here is not restricted in any way to the type of Information System.

In order to explain the provenance algorithm, we next present an example describing the inter-dependence among constituent sources required to integrate a data source named `shipping`. The object `shipping` is described in Appendix B.1.2, as a query called Shipping Priority.

Notation of Figure 4.1: Databases and tables have been denoted by can shapes. Attributes are represented by rounded rectangles, and fusion or transformation functions are shown thorugh oval shapes. The arcs denote the inter-dependence among the sources.

Figure 4.1 shows the integration of `shipping` from three databases called `TPCH`, `TPCHA`, and `TPCHB`.Such databases contain the tables `NATION`, `supplier`,

customers, Orders, and lineitem. The object shipping is composed by the attributes CUST_NATION, SUPP_NATION, L_YEAR, and VOLUME.

Using the entities *DataSourceInfo*, *Ancestors*, and *How_description* the corresponding annotations for the data sources are shown in Tables 4.1, 4.2, and 4.3.

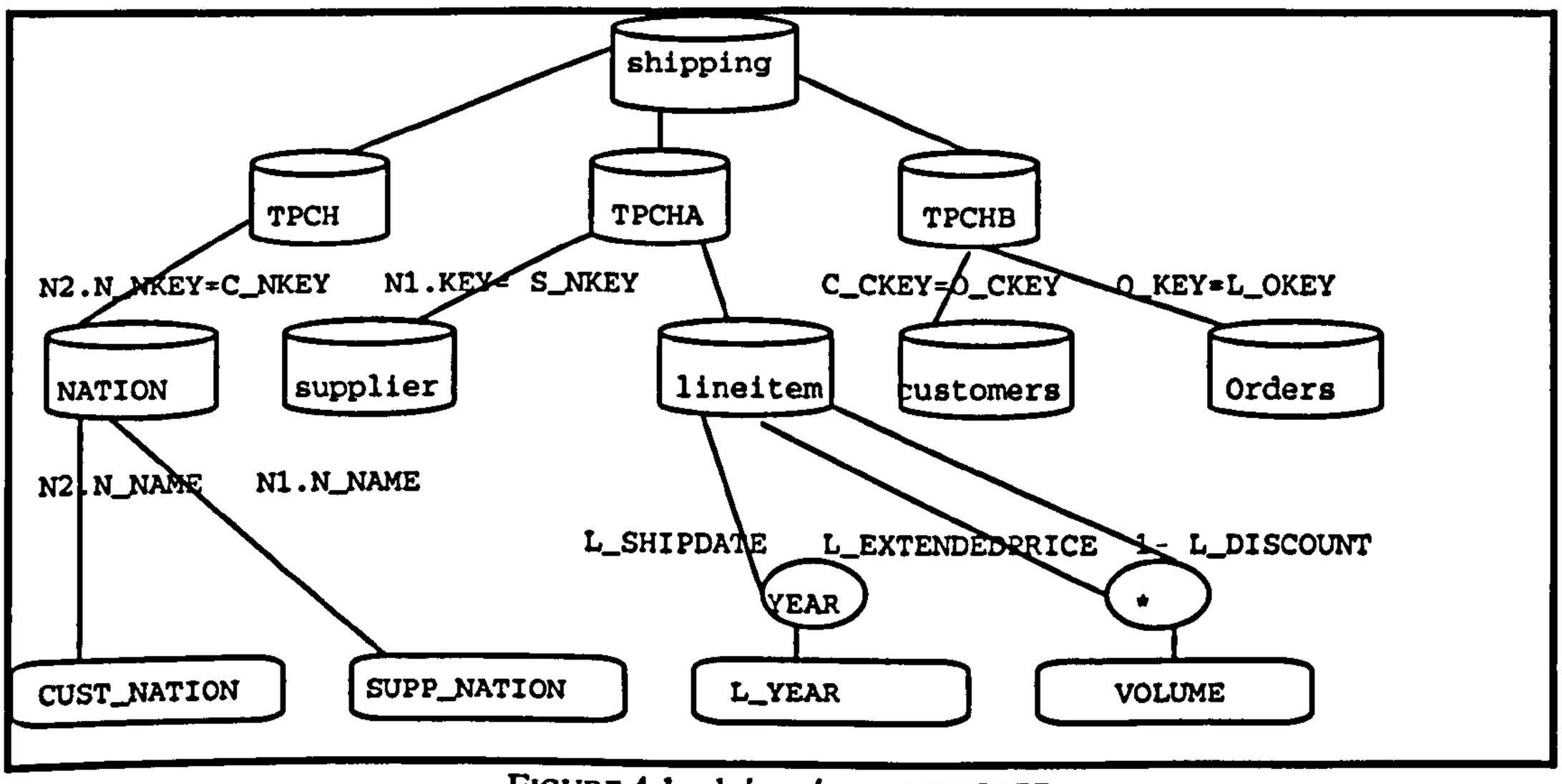


FIGURE 4.1 *shipping* PEDIGREE

Object_id	DataSourceType	PhysicalLocation	object_name	how_ stored	ancestor_id	transformation
200	Database	ASE	TPCH	1	NULL	NULL
201	Table	TPCH	NATION	4	200	NULL
300	Database	ASE	TPCHA	1	NULL	NULL
304	Table	TPCHA	Supplier	4	300	NULL
308	Table	TPCHA	lineitem	4	300	NULL
400	Database	ASE	TPCHB	1	NULL	NULL
406	Table	TPCHB	customers	4	400	NULL
407	Table	TPCHB	Orders	4	400	NULL
409	Derived data	TPCHB	shipping	6	NULL	SELECT SUPP_NATION

TABLE 4.1 *DataSourceInfo*

object_id	ancestor_id
409	201
409	304
409	406
409	407
409	308

Table 4.2 *Ancestors*

How_stored	Description
1	Data entered by DBA as author
2	Data input from the author
3	Data converted by DBA
4	Data come from a Third Party
5	Data converted by Third Party
6	Data fused from more than one source

TABLE 4.3 *How_description*

4.5.1 The provenance algorithm

The following algorithm is not particularly new. We developed it to obtain sufficient information quality from the ancestors to help assessing data quality of derived data (objective 4 of Section 1.5).

The provenance of data is traced by executing queries over the metadata only when it is needed, taking the "lazy approach". The *where provenance* is estimated querying the annotations in the metadata through a recursive algorithm. The tracking stops when it finds the original data source, from all the possible ancestors. Table 4.4 shows the result of the *where provenance* for the data source `shipping`, the table consists of the names of the sources where data is copied from.

Id	Name	Data acquisition method	Ancestor name
409	shipping	Data fused from more than one data sources	
201	NATION	Data comes from a Third Party	TPCH
200	TPCH	Data entered by DBA as author	
304	Supplier	Data comes from a Third Party	TPCHA
300	TPCHA	Data entered by DBA as author	
406	customers	Data comes from a Third Party	TPCHB
400	TPCHB	Data entered by DBA as author	
407	Orders	Data comes from a Third Party	TPCHB
400	TPCHB	Data entered by DBA as author	
308	lineitem	Data comes from a Third Party	TPCHA
300	TPCHA	Data entered by DBA as author	

TABLE 4.4 *where provenance* OF `shipping`.

The *where provenance* in our approach corresponds to a list of sources that had helped in the creation of the data that, together with the query are sufficient to reconstruct the data source in the output as Buneman et al mentions in [Buneman01], see Figure 4.2. It is not our intention to reconstruct the output, but to obtain enough information to compare one data source against other. Moreover, with provenance, users can be aware of the quality of the original source and the changes on the data before they were stored.

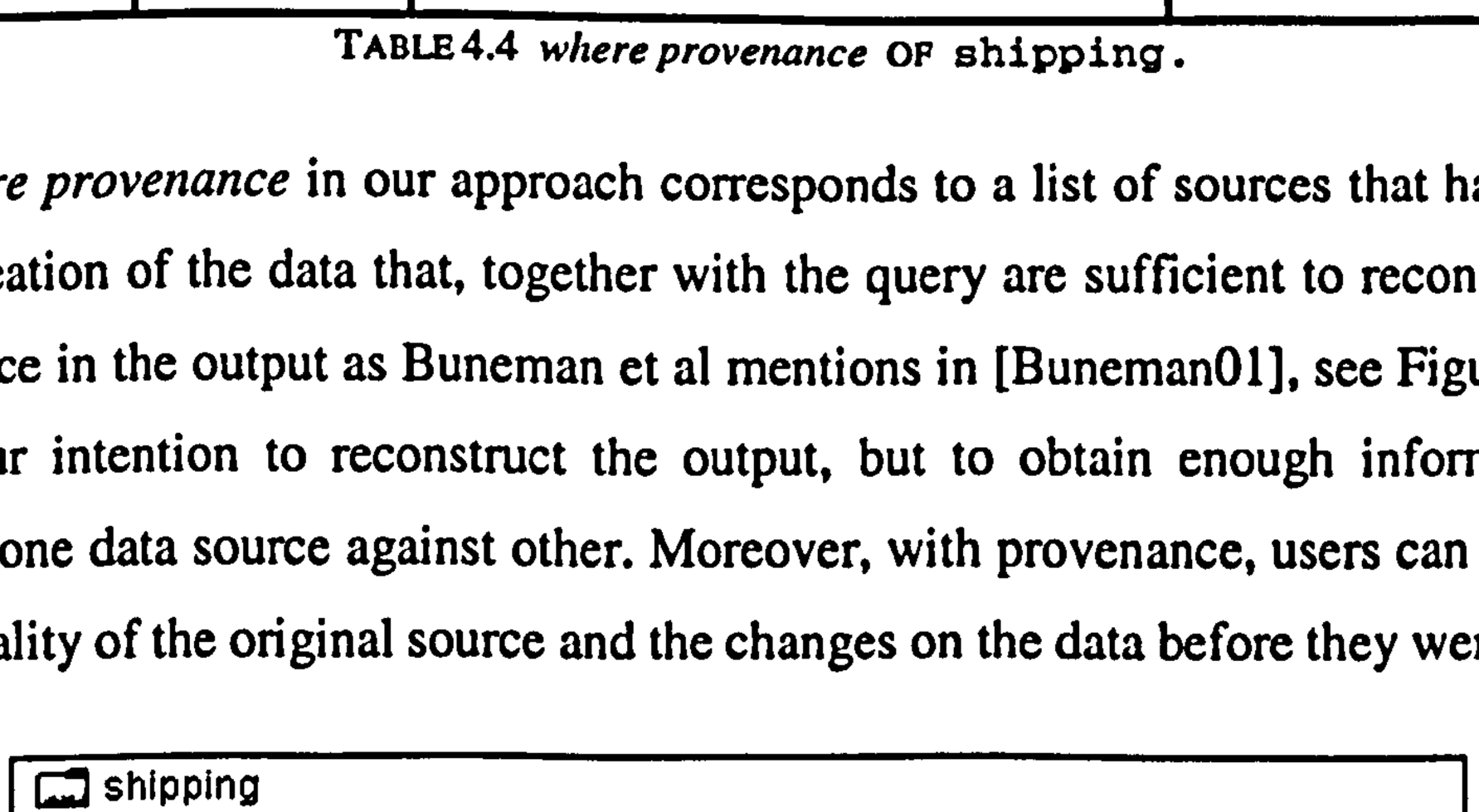


FIGURE 4.2 *where provenance* OF `shipping`

Hence, data provenance is included in the Data Quality Assessment Model because it provides enough information for a better understanding of the data.

4.6 Assessment of derived data based on the quality of its provenance only.

Having explained the process of tracking provenance, suppose that `shipping` can be executed from different queries, then extensional inconsistencies are possible. With the description of provenance, users shall be able to trace back the quality properties of any data by selecting each data ancestor. Therefore to have enough information to trust or not to trust the data. In summary, users are able to compare data sources, and to trust one data source against another by comparing the quality properties of their ancestors.

The provenance algorithm does not trace the queries. So in the case of an attribute derived from a fusion of attributes, and users have no idea which attribute come from which data source, the description of provenance against the attributes of interest is required.

For example we have obtained the provenance of `shipping`, whose *where provenance* is been given by the query and its ancestors as described in the previous section. However, we may be interested in the provenance of its attribute VOLUME which value corresponds to the gross discount revenue. The attribute VOLUME is computed by the formula (VOLUME=L_EXTENDEDPRICE * (1-L_DISCOUNT)), which can be observed from the query shown in Figure 4.2. The *where provenance* is implicit in the query. However, we could ask where L_EXTENDEDPRICE and L_DISCOUNT come from. In the case of extensional inconsistencies, suppose that the attributes of interest can be obtained from two data sources called LINEITEM and `lineitem` as shown in Figure 4.3.

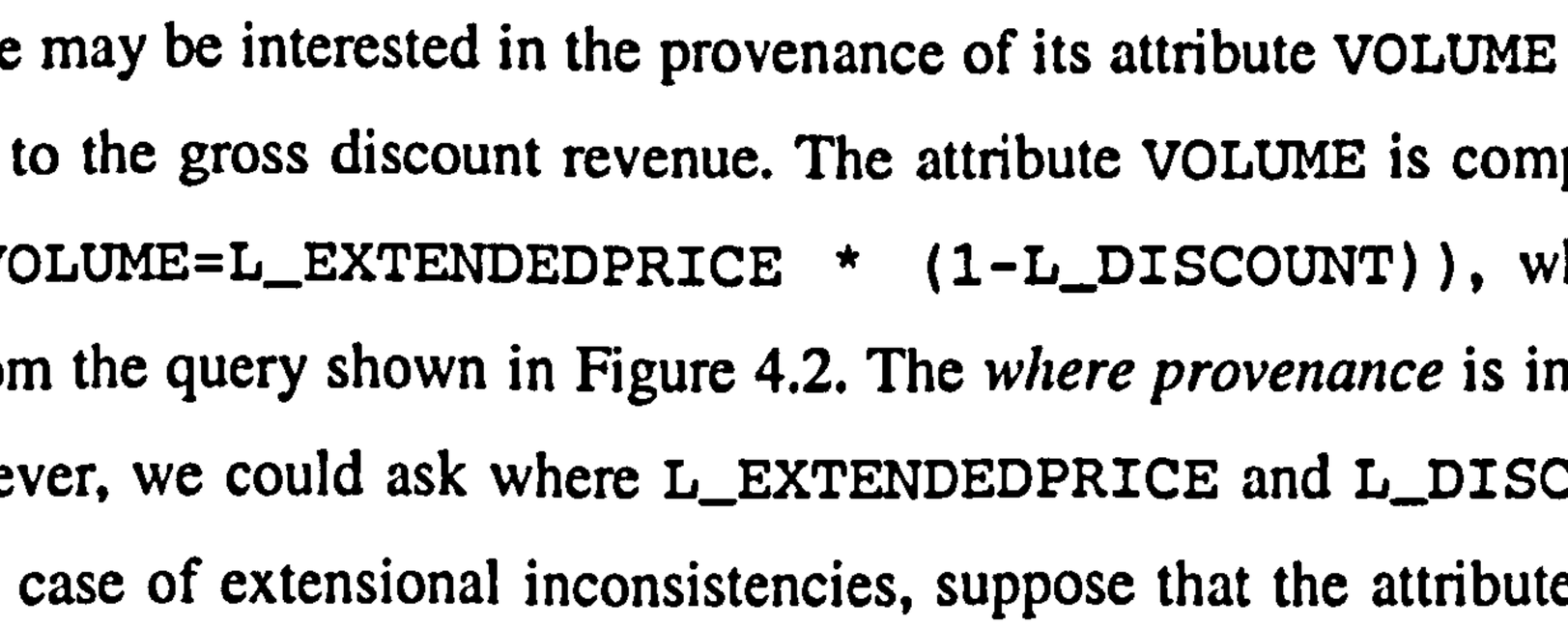


FIGURE 4.3 L_EXTENDEDPRICE AND L_DISCOUNT PROVENANCE

Hence, an analysis of quality of the VOLUME ancestors shall be done in order to determine the quality of the derived attribute. For instance, suppose that LINEITEM is

more accurate and timely than `lineitem`, but less complete. Such criteria should be taken into account to establish an informed decision regarding which query to execute to obtain the gross discount revenue.

4.7 Assessment of derived data by the aggregation of quality of its provenance

We have identified a set of metrics from previous research in section 3.5, refer to [Motro98], [Pipino02], [Naumann00], [Scannapieco04b] for further detail.

We have extended those metrics to assess the quality of primary data sources at multiple levels of granularity to provide qualitative information to users.

Once obtained the assessment of data quality at database, relation, tuple, and attribute levels of primary data sources, the following step was the tracking of data provenance detailed in the previous section.

The data provenance algorithm obtains information relative to the ancestors from which derived data was produced. Once obtained the ancestors, their corresponding quality scores are retrieved to estimate the quality scores of the derived data.

In this case, the DQM is able to assign quality scores to derived data by the aggregation of the quality properties of its ancestors. This assessment requires that all the quality scores of the corresponding ancestors are available.

As we mentioned before, in order to fully achieve our objective 4, we require to aggregate the corresponding scores of the ancestors obtained. In the following examples, we are using the aggregation methods average, and maximum.

We have considered average as a conservative aggregation function for accuracy, completeness, consistency, and uniqueness because we require just an approximation of the quality score. An example is explained as follows:

We are concerned with the accuracy estimation of a data source A. Such data source A is a product of data fusion among three data sources with the following scores of accuracy (X=0.90, Y=0.80, Z=0.60). On the one hand, if we take an optimistic approach then the score would be equal to the maximum value, because it is a positive property. Therefore, A is 0.90 accurate. On the other hand, if we take the pessimist approach, the accuracy score of A is 0.60. However, both approaches are not considering other

implicit elements that might change the score. In the former case, A is also fused from data 0.30 less accurate than the maximum value. In the latter case, the score of A has been decreased from the actual value it may have. The accuracy score of A is not 0.60 neither 0.90, but a mix of the three of them. Therefore, fairer estimation would be 0.76 of accuracy, which is the average aggregation function.

From our research perspective, we believe that average as aggregation function is appropriate for comparison purposes. Furthermore, the same approach should be taken in all fused data situations for consistency. As another example, in the case of time related properties, we could take a pessimistic approach to have an idea of the oldest data value. Otherwise, users would consider data more current than it actually is.

As the main intention of these approximations is to compare data sources, the option will be to take the most current of the data sources fused. Therefore, the maximum function is used for aggregating scores. For instance, we are concerned with the currency estimation of a data source A. Such data source A is a product of data fusion among three data sources with the following scores of currency in days (X=3, Y= 6, Z=2), the currency score will be 6 days. If we compare source A with a currency of 6 days against source B with a currency of 2 days, then the best option would be B because it was compound by at the most 2 days of currency from the oldest component.

For illustrative purposes, in the first case we use average, and in the second case we use maximum as aggregation functions. The rationale is simple the fact that whichever you take as long as all the data will be treated in the same way, there is consistency. The key point is the consistency in the application of the aggregation functions. The formula utilised for fusing data has not been considered in this research. However, the fusion function could be applied for the aggregation of quality scores. Such assessment shall be subject to future research as discussed in Section 4.7.9.

There is no intention in this research to find the optimal aggregation relative to any particular fusion function, but it is something that worths further research.

It is important to mention that the time related quality properties have been measured at a specific granularity level, because it will depend on the Database Management System (DBMS). Such DBMS may provide update time at row level or relation level of granularity. The timestamp obtained at a lower granularity would be preferred against a

timestamp obtained at a higher level of granularity, because it implies the precision of the measure computed due to the capability of the DBMS. These measures should be proportionally scaled for a better precision. Alternatively, we have assumed the worst case as the most common. For instance, if these measures were calculated at the relation level, assume the same time at the tuple, and attribute levels.

In order to explain how quality of derived data might be assessed through data provenance, consider a query or a source s that comes from n ancestors α_j. In the following sections, each metric for primary sources has been already explained in Chapter 3.

4.7.1 Accuracy

Accuracy of derived data $A(s)$ is computed by the average of the scores of its ancestors.

$$Accuracy \qquad A(s) = \frac{\sum_{j=1}^{n} A(\alpha_j)}{n}$$

4.7.2 Completeness

Completeness of derived data $C(s)$ is determined by the average value of the completeness of its ancestors.

$$Completeness \quad C(s) = \frac{\sum_{j=1}^{n} C(\alpha_j)}{n}$$

4.7.3 Consistency

Consistency of derived data $Cn(s)$ is determined by the average of the consistency of its ancestors. The consistency of its foreign keys is checked with its corresponding primary keys in each ancestor.

$$Consistency \qquad Cn(s) = \frac{\sum_{j=1}^{n} Cn(\alpha_j)}{n}$$

4.7.4 Currency

The currency of derived data $Cu(s)$ is the greatest value of the corresponding currency measures from the different ancestors.

Currency $Cu(s) = \max\left(Cu(\alpha_j)\right) \quad , \quad \forall j \in [1 \cdots n]$

4.7.5 Volatility

Volatility is the update frequency. When there are a number of data sources with different volatilities, the volatility of derived data $Vo(s)$ is the greatest value of the corresponding volatility measure from its different ancestors

$Vo(s) = \max\left(Vo(\alpha_j)\right) \quad , \quad \forall j \in [1 \cdots n]$

4.7.6 Uniqueness

Uniqueness of derived data $U(s)$ is obtained from the average of its ancestor's uniqueness.

Uniqueness $U(s) = \dfrac{\sum_{j=1}^{n} U(\alpha_j)}{n}$

4.7.7 Timeliness

Timeliness of derived data $T(s)$ is estimated in terms of its maximum currency and volatility.

$T(s) = \max\left(0, 1 - \dfrac{Cu(\alpha_j)}{Vo(\alpha_j)}\right) \quad , \quad \forall j \in [1 \cdots n]$

4.7.8 Types of assessment

In Section 3.6.2, we have identified two types of assessment associated with the level of granularity of data sources:

Direct assessment: The process of assessment relates directly to the level of granularity such as uniqueness, which relates at the tuple level.

Indirect assessment: The score is calculated based on other scores at other levels of granularity of the same source, such as accuracy at the relation level which value depends on accuracy at the row level.

Having assessed data sources on the bases of its ancestors we are in the position to identify a third assessment method:

Assessment by provenance: The score of an object is computed based on the quality indicators of its ancestors. This assessment method is a novelty of the model and has been designed and will be implemented and tested to prove that data provenance can help to assess data at a finer granularity.

The above granularity-based assessment classification is new, and it is the result of our intention to provide users with a variety of assessments at multiple levels of granularity in order to let users to choose which granularity they are interested in.

4.7.9 Assessment of fused data

We have discussed the assessment of primary data sources in Section 3.5, the assessment of derived data based on the quality properties of its ancestors in section 4.6, and the assessment of derived data based on the aggregation of the quality properties of its ancestors in section 4.7. However, qualitative information is affected when derived data has been obtained from a fusion function. As the provenance algorithm does not trace queries, the algorithm is not able to obtained the elements involved in the fusion function. Therefore the algorithm is not able to aggregate its quality scores and obtain the quality score of fused data automatically. Consequently, the alternative is to assess fused data just in terms of the quality properties of its fused elements. The fused elements can be obtained from the where provenance algorithm through the query that produces the derived data. Such case has been already explained in section 4.6.

We can conclude that the fusion function is not critical because at this point the DQM can still assign a quality value to derived data and specially if it is able to use provenance to track down the fused elements.

However, the fusion function is important for improving the accuracy of the qualitative information that we are providing relative to data elements, and shall be contemplated as part of future work.

The key element for the aggregation of quality scores is to use it consistently. Further research is required for the optimal aggregation within data fusion.

4.8 Summary

Due to the huge amount of heterogeneous data sources available to the user, we can no longer rely on the presumptions of perfection, primary authorship and atomicity.

Therefore, as our approach is also concerned with the quality of integrated data, we have included data provenance as a relevant mechanism to consider in the analysis of data quality.

Woodruff et al and Cui et al have already approached tracking provenance at relational databases [Woodruff97] and [Cui2000].

The algorithm applies tracing queries recursively through the provenance metadata by a lazy approach. This approach is not considering versions or history of data but obtaining enough information to compare one data source against other. The provenance algorithm is not particularly new, it was implemented just to obtain enough information for assessment purposes.

We have considered data provenance as a mechanism to help data quality assessment and consequently, to help in the resolution of data inconsistencies between successor databases. We have proposed the extraction of data provenance using a metadata shared by the community, to demonstrate that provenance is as a helpful mechanism to support the determination of data quality. Furthermore, it is possible to trust data according to the quality scores of its ancestors, or to compute the quality of derived data, considering the scores of its ancestors.

In general, it is easier to trust data obtained directly from its author than data that comes from a number of sources through several transformations. But once the quality scores have been obtained for each data source of interest, this process becomes more precise and informed. Consequently, we can conclude the following:

- The retrieval of the locations of the sources from where the data was extracted, help in the process of data quality assessment.
- The understanding of which data had some influence on the existence of

integrated data helps in the process of data quality assessment.

- Now we are in the position to trace back the originations of data and to obtain quality scores of derived data. Quality of data derived from different data sources at different levels of granularity using data provenance has not been addressed until now.

- Assessment at a lower level of granularity will result in a more precise quality approximation than considering just quality criteria at higher level of granularity.

The Data Quality Manager through the Assessment Model can determine the scores of each participant data source at different levels of granularity. Such scores will be stored in the Data Quality Metadata.

The Measurement and Assessment Models detailed in Chapters 3 and 4 correspond to objectives 2, 3 and 4. These models will be tested through a set of experiments on the prototype in Chapter 7.

Considering the issue of selecting data sources based on customer priorities in order to obtain the best outcome it is necessary to consider the ranking of those data sources based on their quality scores, which we can now compute. Since taking decisions in multiple criteria with values at different units of measure is not an easy task, the next chapter will address the problem.

Chapter 5 Multiple Attribute Decision Making

5.1 Introduction

"Multiple attribute decision making (MADM) procedures, a process for making preference decisions over the available alternatives which are characterized by multiple (usually conflicting) attributes, are useful for improving decision making in a wide range of circumstances" [Hwang95].

We have identified a set of useful data quality criteria for an objective assessment to support a continuous range from expert users to naive users. We have added a provenance mechanism to assess the quality of the participant data sources at different levels of granularity. Now we need to obtain an overall quality measure for each data source according with the quality properties and priorities required by users.

The aim of this chapter is to clarify how the Multiple Attribute Decision Making (MADM) area (developed by Hwang and Yoon in 1981 [Hwang81]) makes possible the ranking of data sources through the comparison of multiple data quality dimensions, different weights, and user quality priorities.

This chapter discusses the MADM problem in terms of data quality criteria, its scores and priorities. We next discuss the scaling criteria methods, the specification of quality priorities by weighting normalization, a brief analysis and evaluation of the available ranking methods. The chapter concludes with some recommendations regarding which ranking method to use according to the scaled scores.

5.2 The Multiple Quality Criteria Problem Definition

Once the quality scores have been obtained, decisions must be made based on the relative quality criteria of data sources.

Consider Table 5.1 with three data sources named TPCH, TPCHA, and TPCHB. There are seven quality criteria per each data source, assigned arbitrarily and their corresponding scores per data source. For instance, accuracy is a positive criterion given

in terms of number of rows, and response time and currency are related to time measures.

	Accuracy	Completeness	Currency	Response Time	Uniqueness	Consistency	Volatility
TPCH	1	0.931	6	2	0.812	1	167
TPCHA	0.611	0.992	24	1	0.484	0.984	150
TPCHB	0.900	0.182	12	0	0.713	0.999	162

TABLE 5.1 M: ORIGINAL SCORES OF TPCH, TPCHA AND TPCHB.

- The first problem to cope with is how to compare multiple and conflicting criteria. Section 5.3 will explain two scaling criteria methods to solve the problems of different units of scores and different ranges.

The table 5.2 shows an arbitrary relative importance for each quality.

Accuracy	Completeness	Currency	Response Time	Uniqueness	Consistency	Volatility
100	65	40	30	90	80	50

TABLE 5.2 W: WEIGHTS ASSOCIATED TO QUALITY PROPERTIES

- The second problem is to handle the quantification of preferences and the detection of inconsistencies every time users set their priorities depending on the situation. For instance, accuracy is the most important property with a weight of 100, followed by uniqueness which weight is 90. However, weights need to be comparable values. Section 5.4 will discuss weighting normalization.

- The third problem is how to combine the quality criteria measurements to produce an overall ranking. Such ranking methods will be presented in Section 5.5.

The MADM methods have already been used to rank data sources based on its quality properties in [Naumann02], [Burgess02] and we are introducing them to provide an overall quality score.

In order to cope with the Multi Attribute Decision Making problem, the above specifications might be given in mathematical terms. Refer to [Hwang81] for further detail.

Let be n the number of alternative data sources A, and k the number of quality dimensions Q. Then the decision matrix can be represented by M.

$$M = \begin{array}{c} \\ A_1 \\ A_2 \\ A_n \end{array} \begin{array}{c} \begin{array}{ccc} Q_1 & Q_2 & \cdot \quad Q_k \end{array} \\ \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,k} \\ m_{2,1} & \cdot & \cdot \\ m_{n,1} & \cdot\cdot & m_{n,k} \end{bmatrix} \end{array} \quad i = 1,\cdots,n \quad , \quad j = 1,..,k$$

The set of priorities assigned to each quality dimension is represented by the vector *W*.

$$W = [w_1 \quad w_2 \quad \cdots \quad w_k]$$

The next section will explain two main scaling methods for scaling criteria onto the [0,1] scale discussed in [Lipschutz02], Vector Normalization and Linear Scale Transformation to compare conflicting criteria.

5.3 Scaling Criteria Methods

The scaling of scores allows all criterion scores to be brought into non-dimensional scores within [0, 1], and thus make them comparable as described in [Naumann02]. The matrix of scaled scores is represented by *N*.

An example of proportion between scores is given by looking at the Table 5.1 and considering the negative criterion Response Time. TPCH takes twice as long as TPCHA to answer the customer. The best response time is given by TPCHB, TPCHA, and TPCH in that order. If the scaling method keeps the proportion between scores, the ranking method will reflect such proportion within the overall score for the ranking.

5.3.1 Vector Normalization

This method divides each score by the norm of the criterion vector for all the alternative sources [Lipschutz02]. $n_{ij} = \dfrac{m_{ij}}{\sqrt{\sum_{i=1}^{n} m_{ij}^{\,2}}}$ (5.1)

(where $m_{i1}, m_{i2}, ..., m_{in}$) is the vector of scores for a given data source

The normalized matrix results from applying equation (5.1) to the original scores are shown in Table 5.3.

It is worth mentioning that this method does not distinguish between positive and negative scores. However, it scales the scores proportionally.

Criteria	Accuracy	Completeness	Currency	Response Time	Uniqueness	Consistency	Volatility
TPCH	0.676	0.678	0.218	0.894	0.685	0.58	0.603
TPCHA	0.413	0.722	0.872	0.447	0.409	0.571	0.541
TPCHB	0.609	0.132	0.436	0	0.602	0.579	0.585

TABLE 5.3 N: SCORES SCALED BY VECTOR NORMALIZATION

5.3.2 Linear Scale transformation

The linear scale transformation method takes the maximum score if the criterion is positive, or the minimum score if the criterion is negative by the formulas 5.2 and 5.3 respectively.

$$n_{ij} = \frac{m_{ij} - m_j^{min}}{m_j^{max} - m_j^{min}}, \quad i = 1, \cdots, n \quad for \quad positive \quad criteria \qquad (5.2)$$

$$n_{ij} = \frac{m_j^{max} - m_{ij}}{m_j^{max} - m_j^{min}}, \quad i = 1, \cdots, n \quad for \quad negative \quad criteria \qquad (5.3)$$

The normalized matrix is presented in Table 5.4, and it is the result of applying the equation (5.2) to accuracy, completeness, uniqueness and value consistency scores, and equation (5.3) to currency, response time, and volatility scores.

Criteria	Accuracy	Completeness	Currency	Response Time	Uniqueness	Consistency	Volatility
TPCH	1	0.924	1	0	1	1	0
TPCHA	0	1	0	0.5	0	0	1
TPCHB	0.744	0	0.666	1	0.698	0.941	0.294

TABLE 5.4 N: SCORES SCALED BY LINEAR SCALE TRANSFORMATION

The scaled scores are in the range [0, 1], where the best option obtains the score 1 and the worst option obtains the score 0. This property assures comparability of scores across criteria. The disadvantage is that if an original score is twice as high as another, this proportion is lost after linear scaling transformation method.

In other words, there is no longer a response time proportion between TPCH and TPCHA. The negative consideration allows considering TPCHB as the best option and TPCH as the worst option. In the case of accuracy, the best option is TPCH and the worst option is TPCHA.

The decision about which scaling method to use should consider two conditions, first the trade off between proportional scaling; and second, the utilization of positive and negative criteria. The ranking methods are addressed in section 5.5.

5.4 Weighting Normalization

It is our intention that users will be able to define their priorities directly, anytime it is required according to different situations.

In the case of k criteria the set of weights is represented as follows:

The priority of each quality attribute is stated by a set of weights, where the MADM require the sum of the elements of the weighting vector be equal 1. Therefore to achieve:

$$\sum_{j=1}^{k} w_j = 1$$

The weighting values must be normalized as follows:

$$w'_j = \frac{w_j}{\sum_{j=1}^{k} w_j} \quad (5.4)$$

The Table 5.5 shows the normalized weights.

Criteria	Original Weights	Normalized Weights
Accuracy	100	0.219
Completeness	65	0.142
Currency	40	0.087
Response Time	30	0.065
Uniqueness	90	0.197
Value Consistency	80	0.175
Volatility	50	0.109

TABLE 5.5 W': WEIGHTING NORMALIZATION

5.5 Ranking Methods

The Multiple Attribute Decision Making area offers several ranking methods. In the following section we are going to explain the Simple Additive Weighting (SAW) and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) methods developed in [Hwang81], and why they were selected among others.

There are more than ten classical MADM methods. Some are either very lengthy, complex, or give biased ranking [Hwang81], which will not be mentioned here because our aim is to find unbiased and practical ranking methods. Therefore, the four most popular and widely used are considered in this section.

- The Simple Additive Weighting method (SAW) is easy to compute and to understand. It allows user interaction to state preferences between criteria.

- The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method is relatively simple, with the same computational complexity as SAW. However, it takes into consideration the differences between positive and negative criteria.

- The Data Envelopment Analysis (DEA) is one of the most widely used. However, it does not allow user weighting. The result is independent of the user.

- The Analytical Hierarchical Process (AHP) requires pair wise comparison for each attribute and a consistency check. Therefore, its computational complexity is exponential to the number of sources and quality dimensions.

These MADM methods have already been used to rank data sources and queries [Naumann02], information items [Burgess03b], and heterogeneous networks [Zhang03].

As we are interested in establishing the user priorities among quality properties DEA is discarded because its results are user independent.

As the AHP complexity is exponential to the number of sources and quality dimensions. AHP can be discarded for complexity reasons.

Further comparisons among these MADM methods have been done in [Nauman02] and [Zhang03]. Based on these evaluations, we have decided to use the SAW and TOPSIS methods.

The following ranking methods require a scaled decision matrix N either using the vector normalization method with the formula (5.1), or the lineal scale transformation with formulas (5.2) and (5.3), and a normalized vector of weights W through the formula (5.4).

5.5.1 Simple Additive Weighting (SAW)

The overall score of an alternative data source is computed as the weighted sum of all the attribute values. This method comprises the following steps:

1- Apply the user defined weighting to each scaled criterion

$$w_j n_{ij} \quad (5.5)$$

2. - The final score SAW for each alternative data source A_i namely SAW(A_i) is therefore calculated by adding the scores up for each criterion as follows:

$$SAW(A_i) = \sum w_j \cdot n_{ij} \quad (5.6)$$

The results of SAW with Linear Scale method example are shown in Table 5.6

Data source	Rank
TPCH	0.813
TPCHB	0.624
TPCHA	0.285

TABLE 5.6 WN: SAW RANKING

5.5.2 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

This method considers that the best alternative should have the shortest distance from the ideal solution and the farthest from the negative-ideal solution.

Considering the decision matrix, the vector of weights:

1. - The normalized decision matrix is weighted.

$$w_j n_{ij} \qquad (5.7)$$

2. - To determine the ideal solution S_j^+ and the negative solution S_j^-

$$S_j^+ = \max(w_j n_{ij}) \quad where \quad i = 1,\cdots,n \ for \ positive \ criteria \qquad (5.8)$$

$$S_j^- = \min(w_j n_{ij}) \quad where \quad i = 1,\cdots,n \ for \ negative \ criteria \qquad (5.9)$$

3. - To calculate the Euclidian distance of each alternative from the ideal solution (E_i^+), and the negative ideal solution (E_i^-), the following equations will be required.

$$E_i^+ = \sqrt{\sum_{j=1}^{k} (S_j^+ - w_j n_{ij})^2} \quad i = 1,\cdots,n \ for \ positive \ criteria \qquad (5.10)$$

$$E_i^- = \sqrt{\sum_{j=1}^{k} (w_j n_{ij} - S_j^-)^2} \quad i = 1,\cdots,n \ for \ negative \ criteria \qquad (5.11)$$

4. - To determine the relative closeness C of the i_{th} alternative to the ideal solution.

$$C_i = \frac{E^-}{E_i^+ + E_i^-} \quad i = 1,\cdots,n \qquad (5.12)$$

Every alternative score is compared with the positive and negative ideal solutions. If an alternative itself is the positive ideal solution then C=1. On the other hand, if an alternative itself is the negative ideal solution then C=0. Therefore, the larger value of C_i, the closer to the ideal solution and farther from the negative solution.

The results of TOPSIS with Vector Normalization are shown in Table 5.7.

Data source	Rank
TPCH	0.677
TPCHB	0.504
TPCHA	0.466

TABLE 5.7 WN:TOPSIS RANKING

5.6 SAW vs. TOPSIS

There are three specific characteristics to consider while ranking data sources:

A) Which combination of scaling and ranking methods is appropriate

B) Which ranking method is more sensitive to weighting or scores

C) Which level of experience users have

The ranking result will vary according with the method used, because there is no "true" ranking of sources [Naumann02]. Ranking of data sources using the linear transformation scaling method with TOPSIS and SAW indistinctly has been approach before [Burgess03b], [Naumann02]. However, the combination between scaling and ranking methods indistinctly might provoke conflicts on the overall quality. Such conflicts are related to the comparison between positives and negatives properties depending in the combination of the scaling method with the ranking method.

In the case of linear transformation, involving positive and negative criteria for example, all the score values are represented in terms of being positives (1 for the best and 0 for the worst option, considering positives and negatives). Then when TOPSIS is used, in the case of negative criteria the method considers the minimum value as the best option, so the comparison is exactly the opposite to what it should be. For instance:

If we use a linear scale transformation (refer to Table 5.4 for further detail), the scaled values for negative criteria such as response time will be the following:

Criteria	Response Time
TPCH	0
TPCHA	0.5
TPCHB	1

TABLE 5.8 SCALED VALUES BY LINEAR SCALE TRANSFORMATION

In the case of negative criteria, TOPSIS will use the minimum score for the best option as already mentioned in equation (5.9). The value chosen for the ideal solution will be exactly the worst option (which is TPCH), degrading the ranking result consequently.

The SAW method will add up the product of weight and the response time score. In the case of TPCH, the result of such product is equal to zero, decreasing the final score. In the case of TPCHB, the result is grater than 0. Such considerations are correct because TPCHB was the best option and TPCH the worst option.

If we use the vector normalization scaling method (see Table 5.3 for further detail), the scaled values for response time will be as follows:

Criteria	Response Time
TPCH	0.894
TPCHA	0.447
TPCHB	0

TABLE 5.9 SCALED VALUES BY VECTOR NORMALIZATION

Again, TOPSIS will use the minimum score (in this case 0) as the best option from equation 5.9, which is TPCHB.

SAW on the other hand, will add 0 to the overall quality score for TPCHB even when it has the best response time. Furthermore, will add a non zero value to the final score to TPCH which actually was the worst option for response time.

In summary, in case positive and negative criteria are involved in the decision matrix. We either use a) Vector normalization scaling method with TOPSIS.

b) Linear scale transformation method with SAW.

If there are just positive criteria in the decision matrix. We use either combination, or

c) Vector normalization scaling method with SAW

d) Linear scale transformation method with TOPSIS

In case, there are just negative criteria in the decision matrix. We use either

e) Linear scale transformation method with SAW.

f) Vector normalization scaling method with TOPSIS.

g) Linear scale transformation method with TOPSIS, and consider all criteria as if they were positive using the formula 5.8 only.

We recommend linear transformation function in the case there are only positive criteria or only negative criteria involved, because this scaling method mantains the proportional differences, and such proportion would make a difference in the outcome.

On the bases of a comparative analysis described in [Naumann02], TOPSIS is more sensitive to weighting and more sensitive to quality criteria with high scores than SAW, which provides a conservative ranking [Zhang03].

As we mentioned before, another aspect for consideration when choosing between SAW and TOPSIS should be user experience. If users have wide experience, the TOPSIS method should be used for a better performance and sensitivity to users preferences. On the contrary, if users are novice they should use the SAW method in order to obtain a conservative ranking or even DEA which does not require any user intervention, but more computational complexity. These considerations shall be taken into account when developing the DQM.

5.7 Summary

This chapter has presented the use of established Multi-Attribute Decision Methods for the ranking of data sources.

We have chosen SAW and TOPSIS methods because they allow user interaction to state preferences between criteria, and because they are easier to compute and to understand than DEA and AHP [Naumann02].

The methods described in this chapter have already utilised for the ranking of data sources by F. Naumann in [Naumann02] and M. Burgess in [Burgess02].

After the use of two Multiple Attribute Decision Making methods, TOPSIS and SAW, it is now possible to rank data sources through the comparison of multiple data quality dimensions, different weights, user quality priorities, at different levels of granularity.

According to our analysis regarding the possible combinations of scaling methods with the ranking methods, the recommendation to obtain coherent results (and therefore a more reliable decision) with both ranking methods, has been to use TOPSIS with Vector Normalization or SAW with Linear Scale Transformation in case positive and negative criteria are involved. In case all criteria are positives or negatives, the Vector Normalization method is the best option with SAW. As far as we know, the previous

analysis has not been done before. In previous approaches such as [Naumann00] and [Burgess02] such distinction was not mentioned.

From our aforementioned outcome, we need to take into consideration two cases: experienced users should choose which ranking and scaling method they prefer. In the case of inexperienced users, the system will use this model to make decisions based on the characteristics of the quality criteria.

Recapitulating: We have discussed our DQ model in Chapter 3, the consideration of Data Provenance in Chapter 4, and the MADM methods to scale, and to rank data sources to produce a single comparable result across data sources (objective 5 of this thesis detailed in Section 1.5) to help users to decide which data source to choose in case of extensional inconsistencies.

Hence, we have all the elements required to carry out the analysis and implementation of a prototype as a proof of concept of our hypothesis (which is the subject of Chapter 6).

Chapter 6 Design and Implementation

6.1 Introduction

The purpose of this chapter is to illustrate the use and implementation of its prototype, which is the objective 6 of this thesis. See Section 1.5.

The DQM has been designed to the solve issues we have found along the research process such as the assessment of derived data, and the integration of data quality scores. The first section of this chapter establishes the specification of the prototype requirements. We then present the analysis and design of the DQM elements, the facilities it provides, and a clarification of its target users. Finally, the chapter presents conclusions regarding the capabilities and limitations of the prototype.

6.2 Requirements

The DQM prototype requirements must correspond to the elements proposed in Chapters 3, 4, and 5. Therefore, the prototype must be capable of supporting the following:

6.2.1 Metadata

The design and implementation of a repository to maintain and to retrieve the quality properties identified in the Data Quality Reference Model called Quality Metadata (QMD). The implementation of a repository named Provenance Metadata (PM) to store the information related to the data sources involved in the multi-database environment, including the data provenance stored in those data sources. Such repository has already been designed in Section 4.4.2.

6.2.2 Data Provenance

Implementation of the algorithm for data provenance from Chapter 4.

6.2.3 Measurement and Assessment Models

This section is concerned with the implementation of the formulas identified in the Measurement and Assessment Models, as follows:

The assessment of primary data sources by direct and indirect assessments as defined in Section 3.6.

The assessment of derived data based on the quality of its provenance as defined in Section 4.6.

The assessment of derived data by the aggregation of quality of its provenance as defined in Section 4.7.

The facility to allow the Data Quality Administrator to execute the above processes each time data changes as part of the temporality of quality properties (as stated in Section 3.6), and to store them in the Quality Metadata.

6.2.4 Ranking of data sources

- The DQM should include the implementation of the Scaling Criteria Methods: a) Vector Normalization; and b) Linear Scale transformation discussed in section 5.3.

- The DQM requires the implementation of the Weighting Normalization method reviewed in section 5.4.

- To carry out the Ranking of data sources, the prototype should compute the methods mentioned in Section 5.5 called a) The Simple Additive Weighting method (SAW) and b) the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS).

6.2.5 Facility to profile the user in terms of the context of the query

Users should be given the ability to prioritise their quality properties and therefore to influence the query outcomes, during the data inconsistencies resolution.

In order to achieve objective 4, the prototype should present the user with a friendly windows interface where they can specify an appropriate context for the analysis of data quality utilising the following options.

- Selection of quality properties

- Specification of quality priorities

- Specification of data sources to analyse
- Specification of scaling and ranking methods

Users should not have to individually select these. There should be a default stereotype condition where scaling and ranking methods will be suggested as was stated in chapter 5. However, experienced users may wish to select and specify all of the conditions for the context of the query.

6.2.6 Analysis of data quality properties

As the DQM should allow users the analysis of the data quality environment, it is important to identify which possible cases may exist within a multi-database environment and the elements it contains.

We have already mentioned that the DQM is a metadata-based system. Therefore, it is possible that for any reason data information may be incomplete or inaccessible. Such information comes from the Provenance Metadata and the Quality Metadata so we have identified 3 conditions:

1. **Analysis of data sources based on their data quality properties only.** There is no ancestor data information at a database level. Therefore, there is no data provenance capability. In such case, users should consider all data in the data source as if they were the primary source.

2. **Analysis of derived data based on the quality properties of its ancestors.** There is metadata stored of the data source and that is sufficient to compute quality scores at a certain level of granularity. In the case that the DQM cannot compute quality properties for derived data, the DQM shall facilitate users the retrieval of data quality by the ranking and analysis of the quality properties of its corresponding ancestors.

3. **Analysis of derived data based on its quality properties.** The quality properties have been computed based not on the idea that data are not the primary source but that we have provenance metadata stored at a data source level that describes the ancestor information of the derived data.

The next section is concerned with the design and implementation of the DQM prototype and presents the facilities developed to support such requirements.

6.3 Design

The DQM is designed according to the requirements under an object oriented programming paradigm. There are three packages namely *Domain*, *User Interface*, and *Database*.

6.3.1 Domain Package

The *Domain* Package contains the class Measurement that is concerned with computing and retrieving the data quality scores by the execution of stored procedures and the scaling and ranking methods.

6.3.2 Database Package

The *Database* Package contains the classes relevant to the Quality Metadata and the Provenance Metadata and the methods required for tracking provenance and retrieving information of the participant databases.

6.3.3 User Interface Package

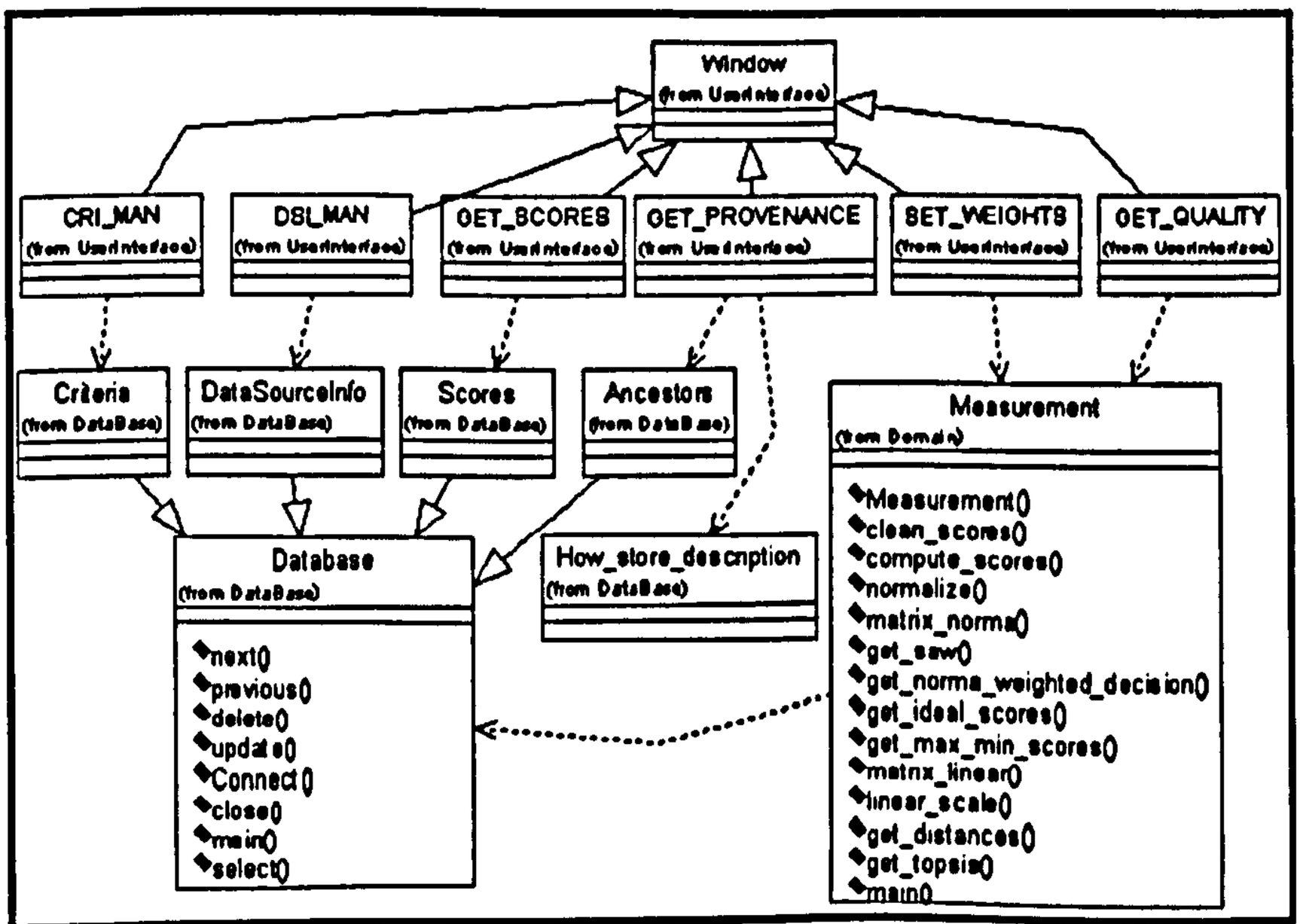


FIGURE 6.1 GENERAL CLASS DIAGRAM OF THE DQM

The *User Interface* Package consists of a number of GUIs that allow the user access to the Reference Model, the tracking of the data provenance, the selection of data sources, the profile specification, and the ranking of data sources according to the options

specified in the main menu. The Class Diagram presented in Figure 6.1 shows very briefly the main classes of the prototype along with their relationships.

6.4 The Data Quality Manager prototype

6.4.1 Prototype Configuration

Operating System

The DQM prototype has been developed and tested on Microsoft windows XP Professional Edition.

Programming Language

The language selected for the development of the DQM prototype was Java Runtime Environment Standard version 1.5.0_05-b05 with Sun Java Development Studio Creator version 6.0. Java was chosen because it is portable, robust and the author is familiar with the language. A Java interface has been provided for a number of enterprise Databases Management Systems, such as those from Oracle and Sybase.

Database Management System

The implementation of the Quality and Provenance Metadata was performed with Sybase Adaptive Server Enterprise version 12.5.2. This Database Management System was selected because the author is well acquainted with the software and the capability to build stored procedures for measurement and assessment.

In order to illustrate the implementation of the DQM the next section will show the most relevant screens according to the requirements established from the options available in the main menu. Such screens can be identified by name from the class diagram in Figure 6.1.

6.4.2 Main menu

The main menu is a tool bar representing each element of the architecture of the Data Quality Manager. Some elements of the tool bar are available depending on the status of the user and therefore for a fully working system a password entry system would precede the stage to ensure that user status was identify. See figure 6.2.

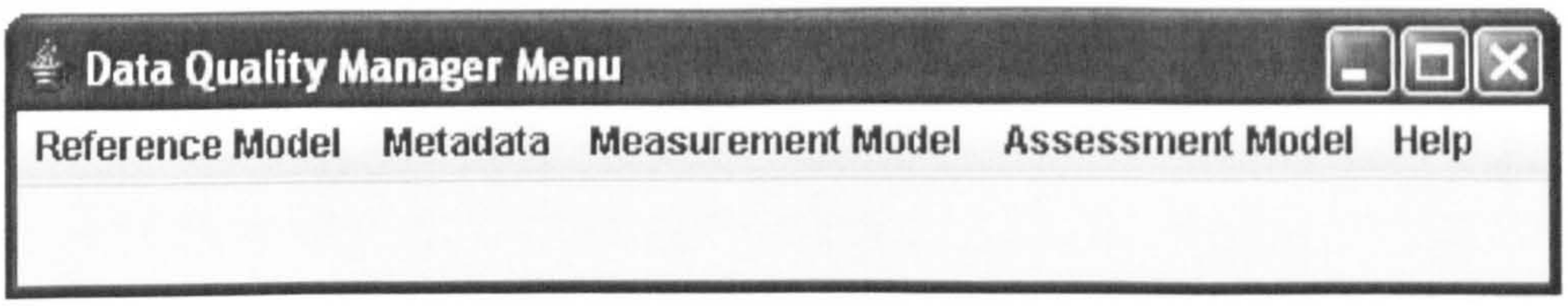


FIGURE 6.2 DATA QUALITY MANAGER MAIN MENU

6.4.3 Reference Model

According to the requirements established in Section 4.1 of this chapter, the first option called *Reference Model* allows insertion, deletion and update of quality properties from the Reference Model, and corresponds to the CRI_MAN window in the diagram class. This menu selection is only available for use by the Data Quality Administrator (DQA).

6.4.4 Metadata

We have implemented the Quality Metadata within relational database architecture. The Entity-Relationship Diagram for the Quality and Provenance Metadata is shown in Figure 6.3.

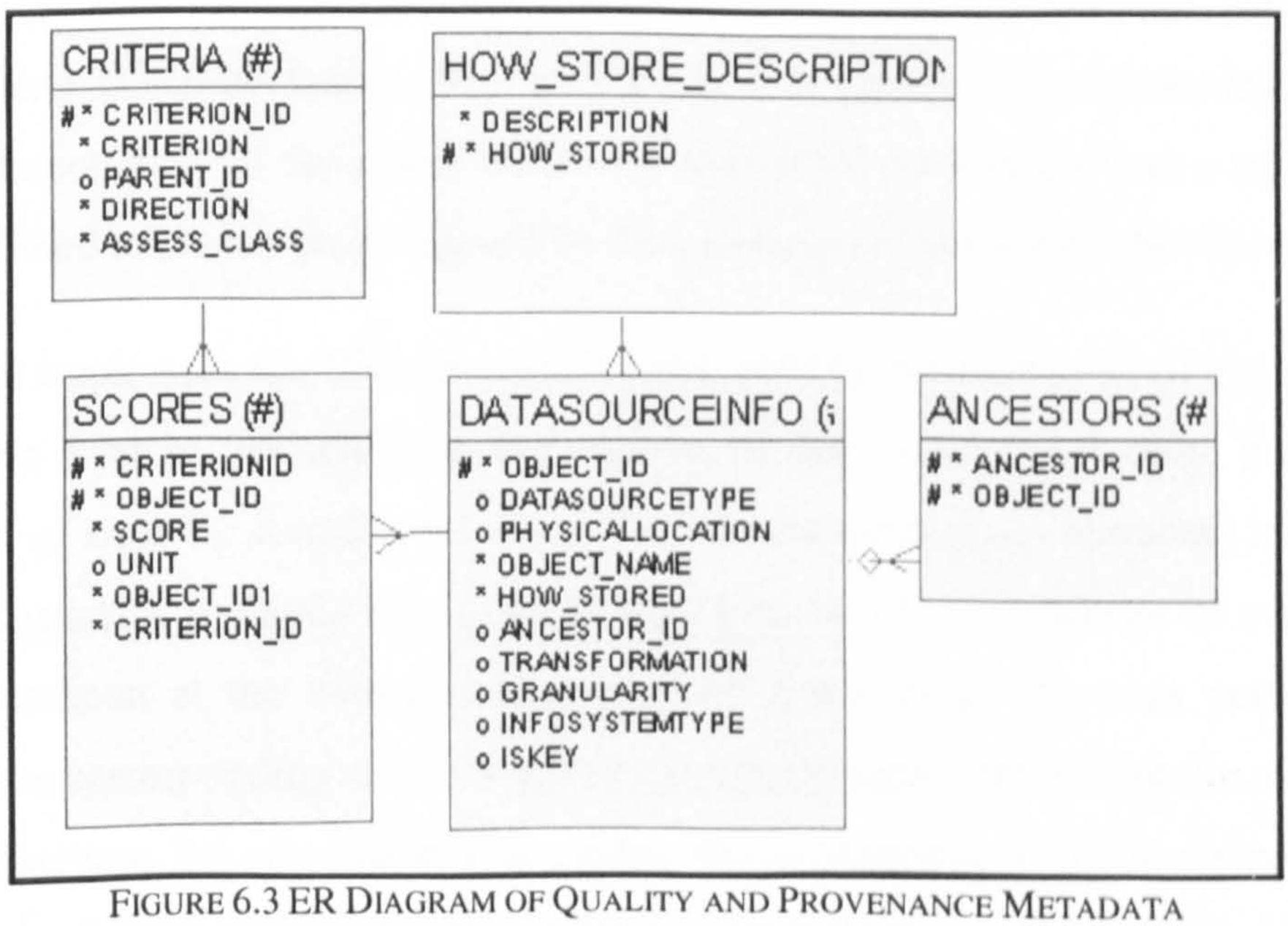


FIGURE 6.3 ER DIAGRAM OF QUALITY AND PROVENANCE METADATA

The *Metadata* option allows the retrieval and management of the information related to the participant data sources, and corresponds to the requirement indicated in Section 6.4.1. This activity is only available to the Data Quality Administrator. See Figure 6.4.

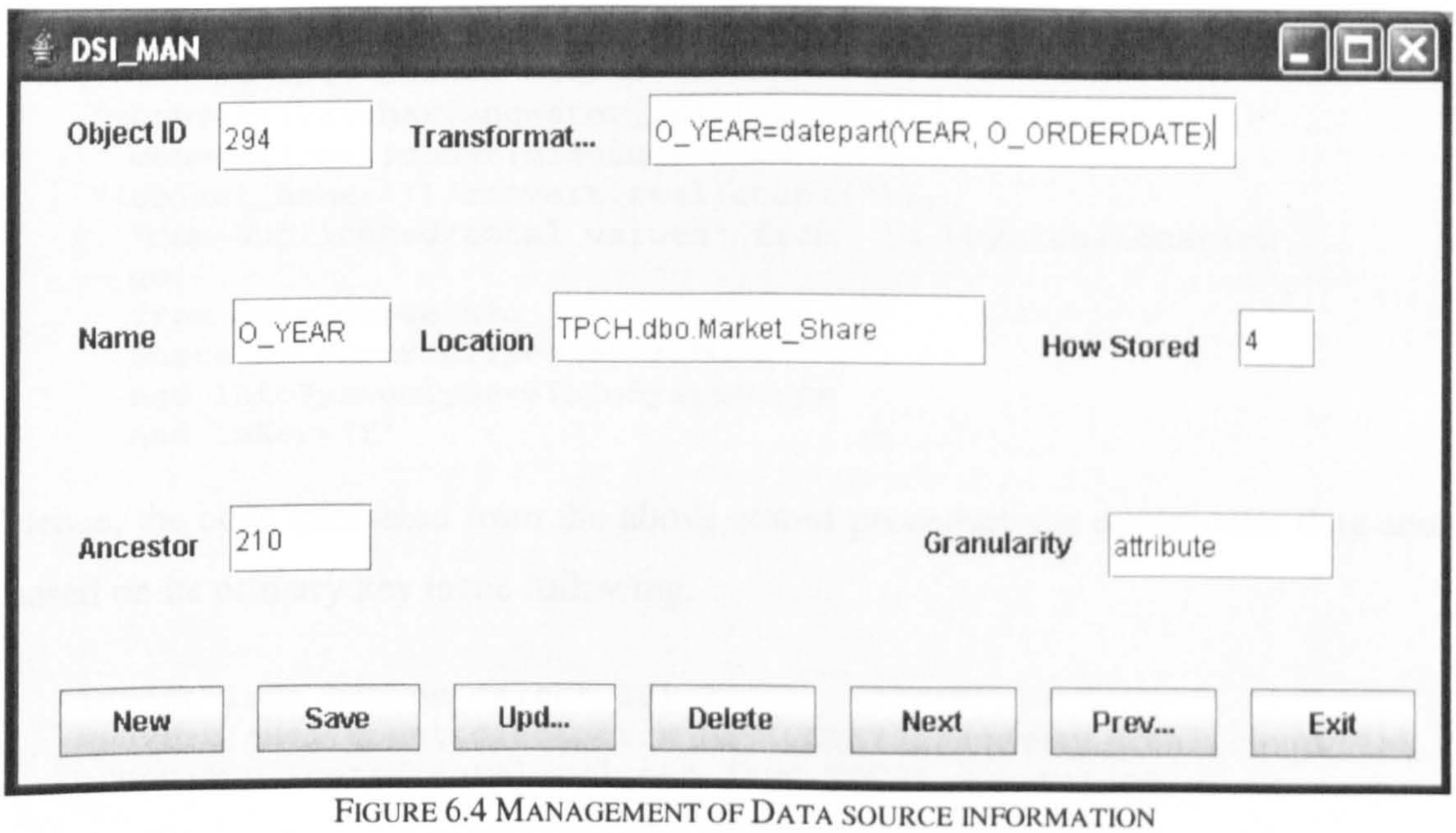


FIGURE 6.4 MANAGEMENT OF DATA SOURCE INFORMATION

6.4.5 Measurement Model

The *Measurement Model option* corresponds to the requirements stated in Section 6.4.3 and it contains two submenus *Compute Scores* and *View Scores*:

The submenu *Compute Scores:* The purpose of this option is to let the Data Quality Manager to compute all the scores within the federation automatically to a certain point to a predefined tolerance period agreed by data consumers as stated in Section 3.6.1.

In order to implement the Measurement Model without consideration of any particular Information System, according to the metrics at data value level only, there was a generation of code by stored procedures. Such stored procedures extracted information from the metadata, to obtain the names of the elements of the databases of every object in the federation at the corresponding level of granularity. This was performed by applying the corresponding metric of every quality property, and storing the result in the Quality Metadata. We consider this as a relevant characteristic of the implementation of the Measurement Model. Therefore, we present the following code as an example of the generation of SQL for Uniqueness metric.

```
  select '
  insert into Scores select
'+convert(varchar,ancestor_id)+',2,
  convert(real,count(distinct
'+object_name+'))/convert(real,count(*)),
  "non-duplicated/total values" from  '+ PhysicalLocation '
  go'
  from DataSourceInfo
  where DataSourceType='attribute'
  and InfoSystemType=@InfoSystemType
  And isKey='Y'
```

Hence, the code generated from the above stored procedure for a particular data source based on its primary key is the following:

```
insert into Scores select 301,2,
convert(real,count(distinct N_NATIONKEY))/convert(real,count(*))
,"non-duplicated/total values" from TPCHA.dbo.nation
```

The above SQL code computes the ratio between the number of non-unique rows and the total number of rows in the `nation` relation.

Having computed the scores from primary data sources (only one level of indirection, direct parent, no ancestor), the next relevant step is the computing of quality scores of derived data, which are determined by the corresponding scores of their ancestors. The next code shows an example for the quality property uniqueness.

```
insert into Scores
select 1209,2,avg(score),"avg(uniqueness of ancestors)"
from Scores
where object_id in (select ancestor_id
                    from Ancestors
                   where object_id = 1209)
and criterionID=2
```

The submenu *View Scores*: This option allows users to retrieve the existing quality scores for a specified data source (see Figure 6.5).

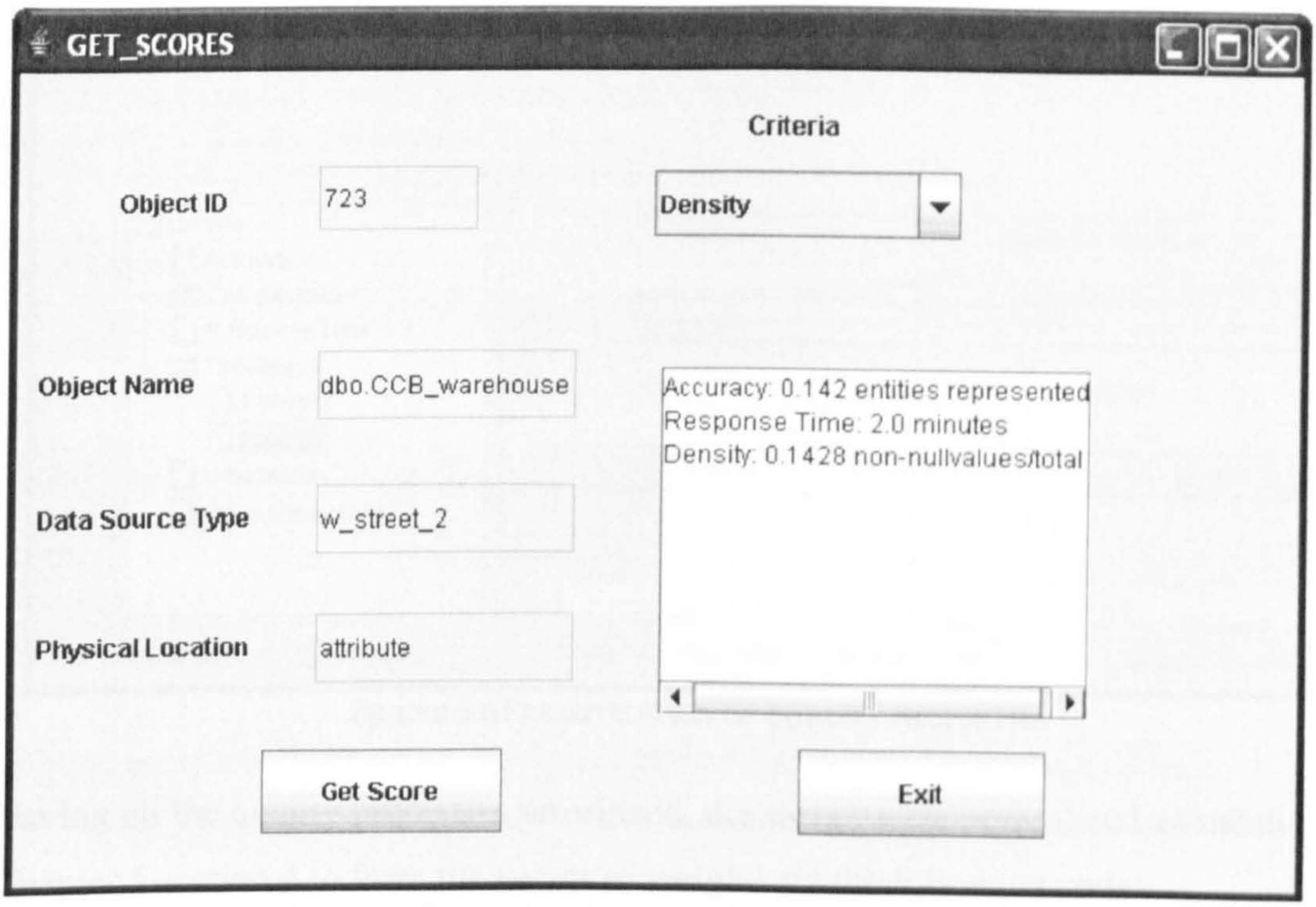


FIGURE 6.5 DISPLAY OF QUALITY SCORES

6.4.6 Assessment Model

The *Assessment Model* option has two submenus *Ranking of Data Sources* and *Provenance* (requirements 4.5 and 4.2 respectively).

The submenu *Ranking of Data Sources:* This option allows the ranking of data sources by the specification of quality properties, quality priorities, and ranking methods.

The facility to profile the user in terms of the context of the query depends on the level of experience of the users, in the case of an experienced user is implemented as follows:

The system displays a tree with all available quality criteria from the quality metadata for selection, and for each specified quality criterion, a slider is displayed to set the corresponding weight. See Figure 6.6.

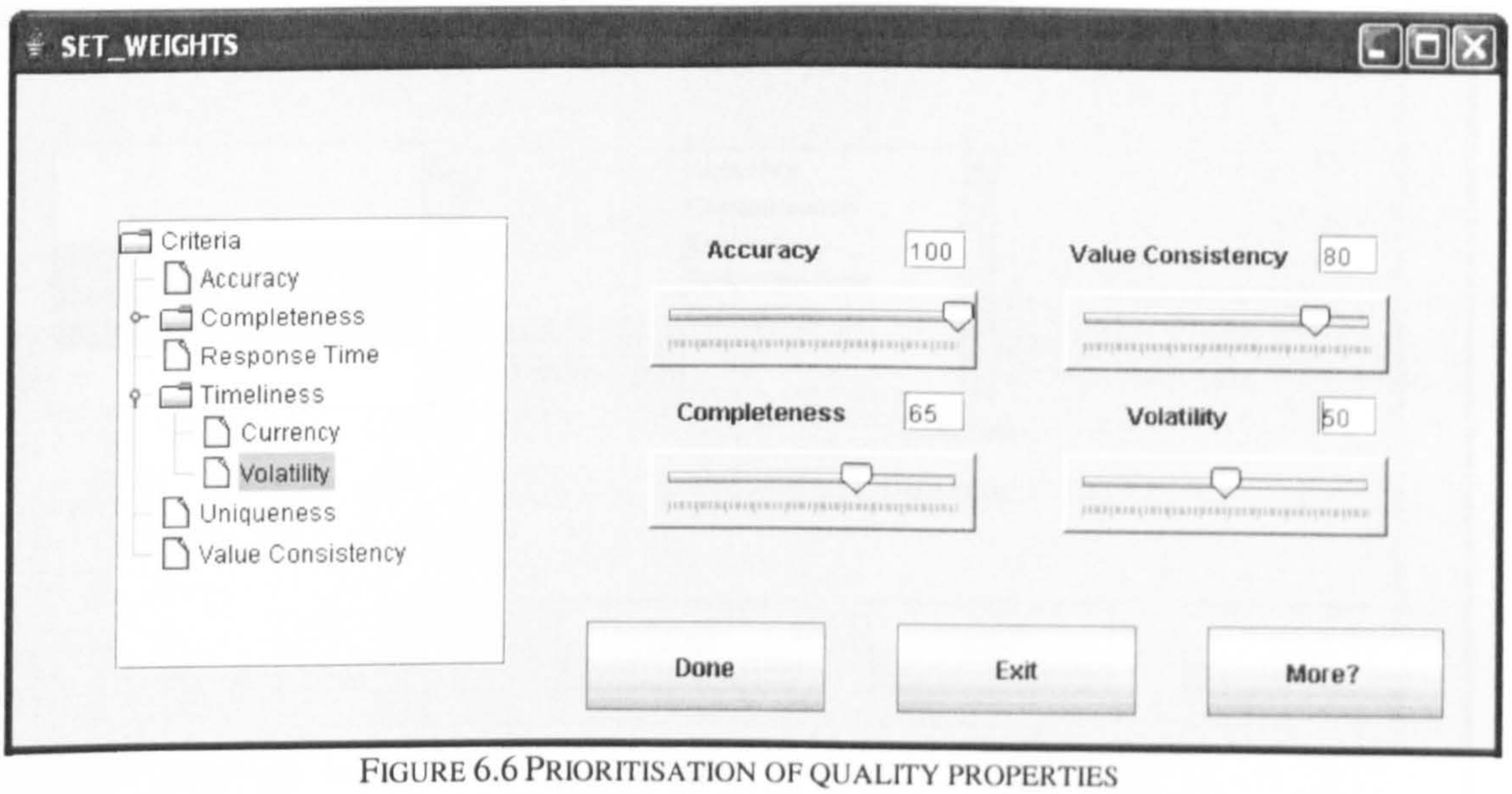


FIGURE 6.6 PRIORITISATION OF QUALITY PROPERTIES

Having all the quality properties prioritised, the weights are normalized as mentioned in Chapter 5 section 4 to form the vector of weights by the following code:

```
public void normalize()
  {  int i,sum=0;
  for (i=0;i<this.chosen_length;i++)
        sum+= this.weights[i];
  for (i=0;i<this.chosen_length;i++)
        this.weights[i]/=sum;
         }
```

The next step is the selection of the data sources, scaling, and ranking methods. In order to select data sources from a scroll pane, the prototype retrieves from the metadata all the data sources involved in the federation of interest (highlighted in Figure 6.7). The scaling method is selected by pressing its corresponding radio button and the ranking of data sources is executed by pressing the buttons TOPSIS or SAW. In the example the linear scale transformation with SAW methods were executed. The following code shows the implementation of the SAW method.

```
public void get_saw()
  {double sum=0.0000000;
   this.saw = new double[num_sources];
   for(int i=0; i<num_sources; i++)
   { for(int k=0; k<chosen_length; k++)

sum+=this.norma_decision[i][k]*this.weights[k];
    this.saw[i]=sum;
    sum=0.00000000;
    }   }
```

The overall quality is presented in descendent order in a Text area. See Figure 6.7.

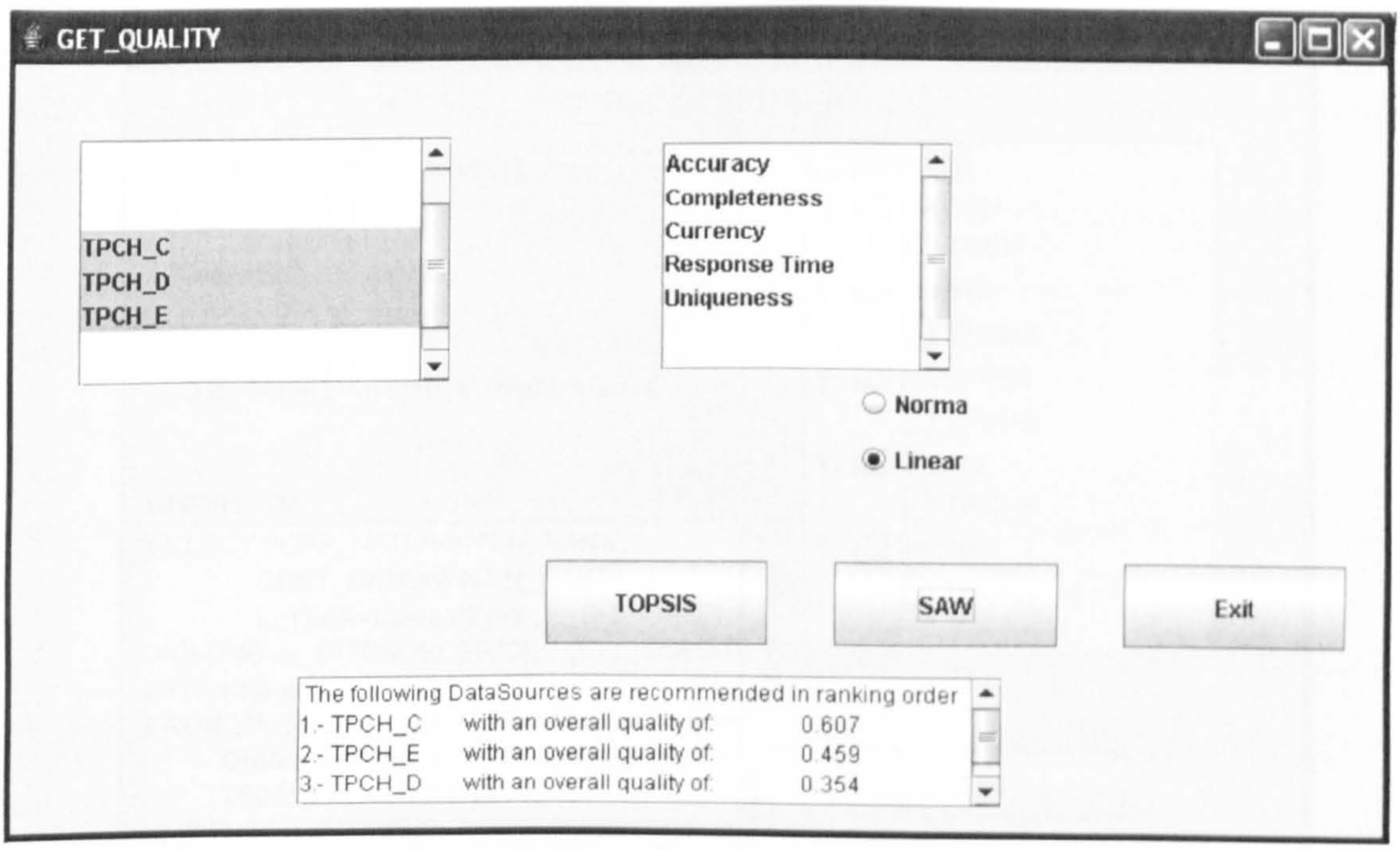


FIGURE 6.7 SELECTIONS OF DATA SOURCES, SCALING AND RANKING METHODS

The submenu *Provenance:* This facility obtains the provenance of a selected data source and the quality scores available. Therefore, users are able to select data sources and their quality properties, and then proceed to the ranking of data sources.

The first step is to specify the name or identifier of the data source whose provenance is required. The Data Quality Manager provides users with a friendly hierarchical tree describing the data lineage, to trace back to sources through data paths. Every element of the tree can be selected and its available quality scores are displayed in a table. (See figure 6.8).

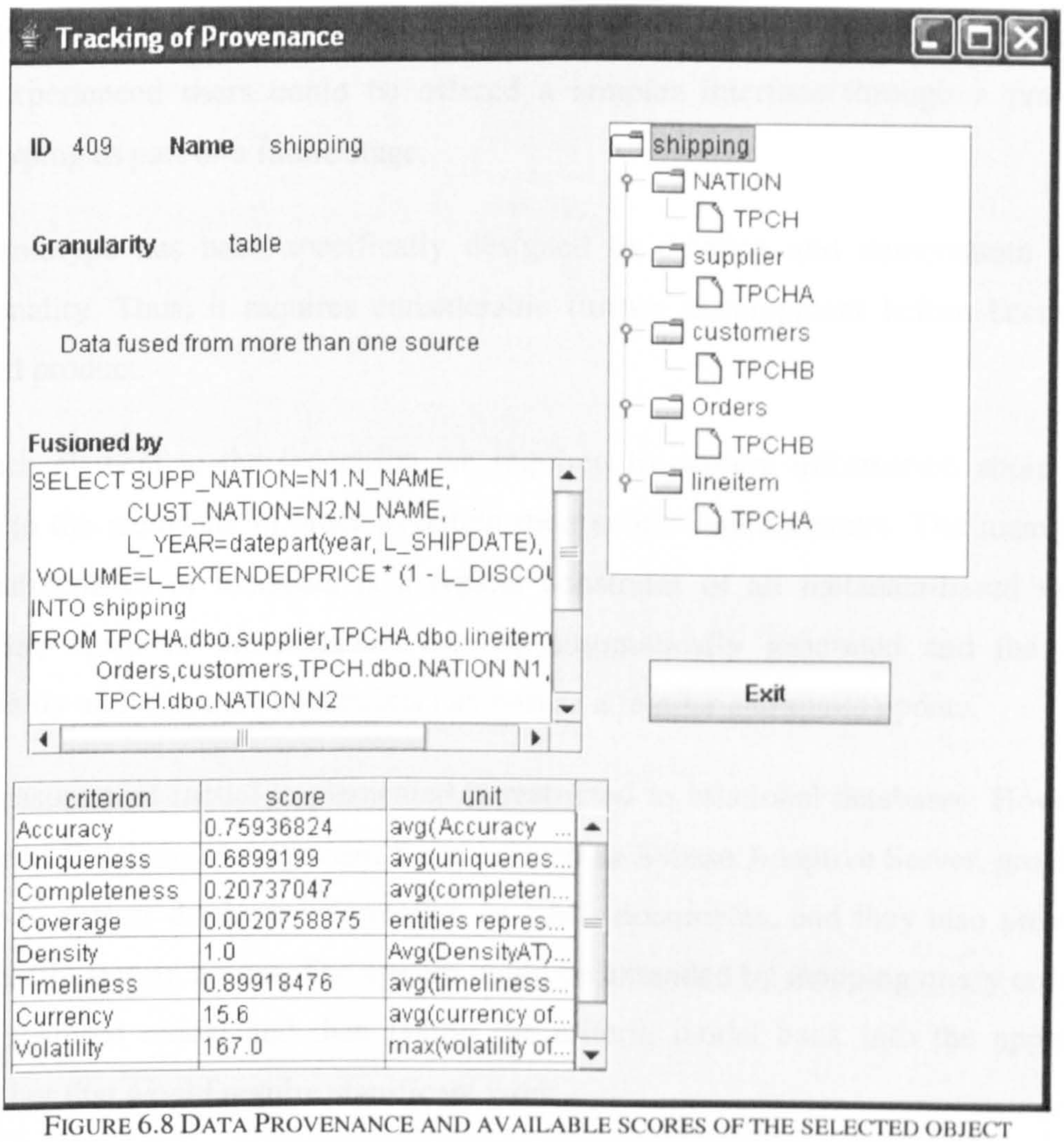


FIGURE 6.8 DATA PROVENANCE AND AVAILABLE SCORES OF THE SELECTED OBJECT

6.5 Conclusions

In order to achieve the expected functionality, we have taken the architecture of the Data Quality Manager and reflected on its requirements (covered in Chapters 3,4,and 5).

We have built a system prototype where we have an understanding of the potential cases that can exist in the database. These possible cases reflect all possible conditions because a combination of them will reflect any database condition. Chapter 7 will present an example of three scenarios representing the cases mentioned in section 6.4.6 in order to test the DQM prototype.

The Data Quality Manager allows users to obtain qualitative information on the basis of a set of quality properties, measurement, and assessment of data quality.

The DQM also presents a mechanism for users to define the data criteria, quality criteria, and then goes to the lowest level of granularity in the ancestor database.

The system supports expert users in allowing them explicitly to operate at this level. Less experienced users could be offered a simpler interface through a process of stereotyping as part of a future stage.

The prototype has been specifically designed to develop and demonstrate specific functionality. Thus, it requires considerable further development before becoming a finished product.

For each element in the federation we required to capture information about a data source in the metadata, in order to obtain their scores and ancestors. The management and maintenance of metadata is a typical constraint of all metadata-based systems. However, much of the metadata can be automatically generated and the use of temporality and tolerance constraints can ensure a regular automatic update.

The measurement model implemented is restricted to relational databases. However, a number of Database Management Systems such as Sybase Adaptive Server, provide the facility to extract database information as XML documents, and they also provide an external file system access. The system could be extended by mapping query constructs into a required model and then taking the criteria model back into the appropriate system but that would require significant work.

As we mentioned in Chapter 3, the implementation of the DQM was restricted to the data value level quality properties. However, it can be easily extended to the measurement of subjective data quality properties, and store these measurements as part of the user profile. Such extension would be helpful in case of experienced users.

This model is not restricted to any type of data source in terms that data provenance has already been tracked for documents and semi-structured data [Buneman01]. In such a case, the assessment of data quality by data provenance still works.

As can be seen from the description of the implementation of the DQM prototype, there are a number of possible combinations of quality properties, quality priorities, and levels of granularity according to the context established by the user that might change the ranking of data sources and therefore the query outcomes.

Having designed and implemented the Data Quality Manager, we are now in the position to test the functionality of the prototype, the appropriateness of the quality

information, the ranking of the data sources, and that the qualitative information varies appropriately according with the context specified by the user by testing and experiments detailed in Chapter 7.

Chapter 7 Test and Experimentation

7.1 Introduction

This chapter presents a test plan and an experimentation plan. Thus, the main goals are:

1. Testing the functionality of the prototype.
2. Testing the appropriateness of the quality information.
3. Testing the ranking of the data sources.
4. Testing the appropriateness of the prototype within the test boundaries.

7.2 Test Plan

Testing involves the establishment of objectives, resources, strategies, test cases and procedures for handling problems. Therefore, in order to test the functionality of the system, we need to identify which relevant functional requirements will need validation through a set of activities. Evaluating how well these activities support the requirements will lead us to the determination of whether the prototype achieves the required functionality or not (covered in Section 7.2.2).

Regarding the appropriateness of the quality information synthesized, we have identified representative scenarios according with the data quality properties assessed on specific data sources to test whether the scores produced by the prototype are within the expected ranges (referred to in section 7.2.3).

Once the data sources have been evaluated, we require testing if the ranking of the data sources produced by the DQM reflect the expected outcomes (covered in Section 7.2.4).

To test the potential capabilities of the DQM prototype to determine if it behaves as it is expected by providing appropriate qualitative information of derived data, and at database, relation, attribute levels of granularity by conducting an experimentation plan based on inferential and descriptive statistics through a set of sanity checks.

7.2.1 Testing Configuration

Operating System

The DQM prototype has been tested on Microsoft Windows XP Professional Edition.

Database Test Suite

As the DQM prototype is aimed to work within a multi-database environment, the conducted tests are based on two populations corresponding to the following benchmarks:

- TPC Benchmark™C (TPC-C) is an on-line transaction processing benchmark. TPC-C simulates a complete computing environment where a population of users execute transactions against a database. The benchmark is concerned with the execution of transactions of an order-entry environment. These transactions include entering and delivering orders, recording payments, checking the status of orders, and monitoring the level of stock at the warehouses [TPC-C]. We have implemented 6 databases which architecture is further detailed in Appendix B. Such databases have been loaded and strategically updated to reflect differences of quality.

- The TPC Benchmark™H (TPC-H) is a decision support benchmark. It consists of a suite of business oriented ad-hoc queries and concurrent data modifications. The queries and the data populating the database have been chosen to have broad industry-wide relevance. This benchmark illustrates decision support systems that examine large volumes of data, execute queries with a high degree of complexity, and give answers to critical business questions [TPCH]. We have implemented 6 databases which architecture is further detailed in Appendix B. Such databases have been loaded and strategically updated to reflect differences of quality.

Database Management System

The TPC benchmarks were implemented within Sybase Adaptive Server Enterprise version 12.5.2.

7.2.2 Testing the functionality of the prototype

We have identified that the analysis of data quality properties is our main functional requirement, so in order to test it; we have mapped the three test cases identified in section 6.4.6. Besides, every condition case contains requirements previously identified in Chapter 6, is also showed in Table 7.1.

Test case 1: Analysis of data sources based on their data quality properties only.

Test case 2: Analysis of derived data based on the quality properties of its ancestors.

Test case 3: Analysis of derived data based on its quality properties.

Test Case	Quality Metadata	Provenance Metadata	Tracking Provenance	Measurement and Assessment	Profile Query context	Ranking of Data Sources
1	√	X	X	√ Direct and indirect	√	√
2	Incomplete	Incomplete	√	√ Direct and indirect	√	√
3	√	√	√	√ With provenance	√	√

TABLE 7.1 OUTLINE OF TEST SUITE

We will conduct the tests by the functional technique approach, based just on input and output, to take an end user testing perspective.

The test suite is designed to consider the requirements of expert users because we can get a high level of detail in the analysis of the data and the functionality of the system. Besides, the level of functionality required for less expert users diminishes because the system will automate it according to the user stereotypes. Thus, the greatest level of customisability of the system is at the level of expert users so that is why we tested at that level.

Test case 1: Analysis of data sources based on their data quality properties only.

As there is no provenance capability as shown in Table 7.1, users consider all data in the data source as if it were the primary source.

Scenario: An experienced user has three databases (TPCH_C, TPCH_D, and TPCH_E) available to query from. In order to identify which data source could bring better data to establish an informed decision, the user has considered that accuracy, uniqueness, consistency, completeness, volatility, currency and response time are the quality properties required (in that order). The context user is summarized in Table 7.2.

Information System	Granularity Level	Ranking Method	Quality Criteria	Weights
TPCH	Database	SAW	Accuracy	100
			Completeness	65
			Currency	40
			Response Time	30
			Uniqueness	90
			Consistency	80
			Volatility	50

TABLE 7.2 CONDITIONS FOR THE ANALYSIS OF DATA QUALITY WHITIN A DSS

Procedure The first step is to select the domain of the Information System. Once selected, the next step is the specification of the desired quality properties along with their corresponding weights see Figure 7.1.

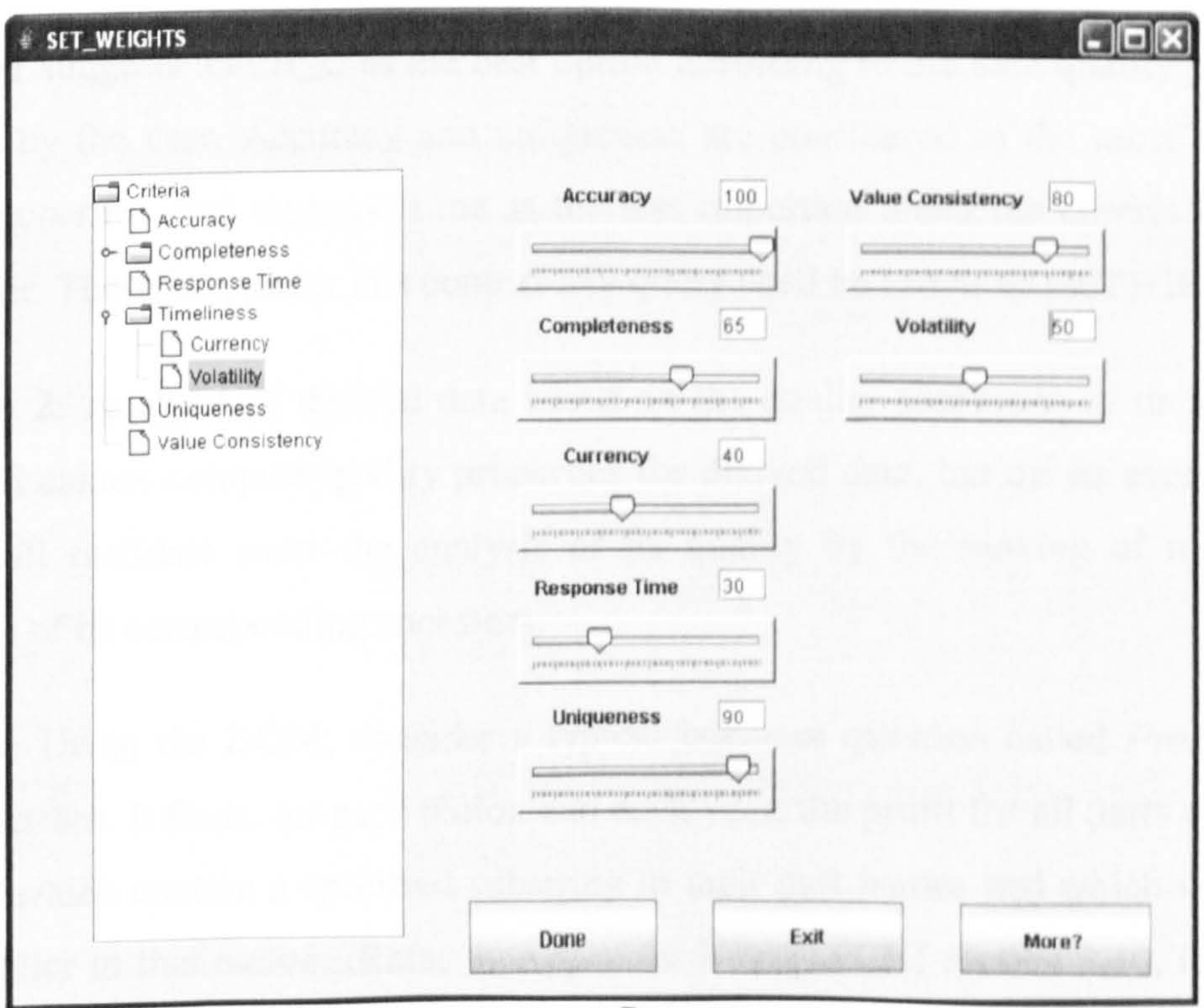


FIGURE 7.1 SETTING OF QUALITY PROPERTIES AND ITS WEIGHTS

In the third step, the user selects TPCH_C, TPCH_D, and TPCH_E as the primary data sources, and linear transformation as the scaling method with the SAW method for the ranking of data sources. From the above scenario, we tested that the user could set some priorities, having identified the important criteria and the secondary criteria. The outcome corresponding to the ranking of data sources is shown in Figure 7.2.

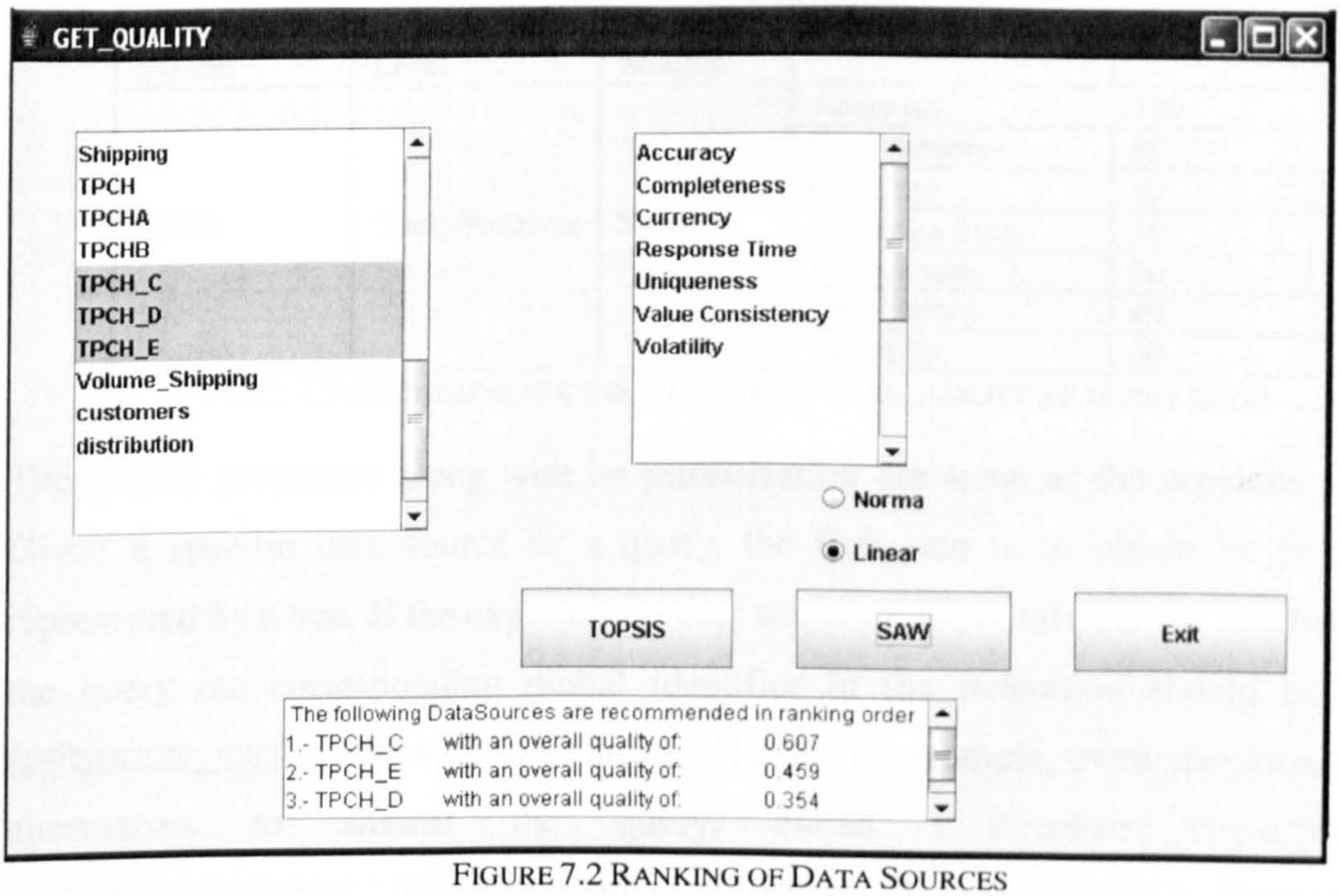


FIGURE 7.2 RANKING OF DATA SOURCES

Results

The DQM suggests TPCH_C as the best option according to the data quality properties specified by the user. Accuracy and uniqueness are considered as the most important quality properties, and response time as the less important under the criteria identified by the user. Therefore, under this context any query shall be executed on TPCH_C.

Test case 2: Analysis of derived data based on the quality properties of its ancestors. The DQM cannot compute quality properties for derived data, but for its ancestors, the DQM shall facilitate users the analysis of its quality by the ranking of the quality properties of its corresponding ancestors.

Scenario: Using the DQM, consider a typical business question called *Product Type Profit Measure*. It finds, for each nation and each year, the profit for all parts ordered in that year which contain a specified substring in their part names and which were filler by a supplier in that nation. (Refer to Appendix B or [TPCH] section 3.10, for further details on the business question).

This scenario takes into consideration the possibility that for any reason the scores of the query *Product Type Profit*, could not been estimated. If this is the case, we can still have an informed decision by using the DQM. Consider the conditions established in Table 7.3.

Information System	Granularity Level	Ranking Method	Quality Criteria	Weights
TPCH	Query/Relation	SAW	Accuracy	100
			Completeness	65
			Currency	40
			Response Time	30
			Uniqueness	90
			Consistency	80
			Volatility	50

TABLE 7.3 CONDITIONS FOR THE ANALYSIS OF DATA QUALITY AT QUERY LEVEL

The quality properties along with its prioritization are same as the previous scenario. Given a specific data source or a query, the first step is to obtain its provenance represented by a tree. If the experienced user wants to obtain all possible alternatives for the query the corresponding global identifier in the federation should be entered (`PRODUCT_TYPE_PROFIT` with an id 511). In the example, there are three possible alternatives to answer the query, called `C_Product_Type_Profit`, `D_Product_Type _Profit`, and `E_Product_Type _Profit`. Every time an alternative is selected from the tree, its corresponding ancestors will be displayed along with the query code of such alternative. See Figure 7.3.

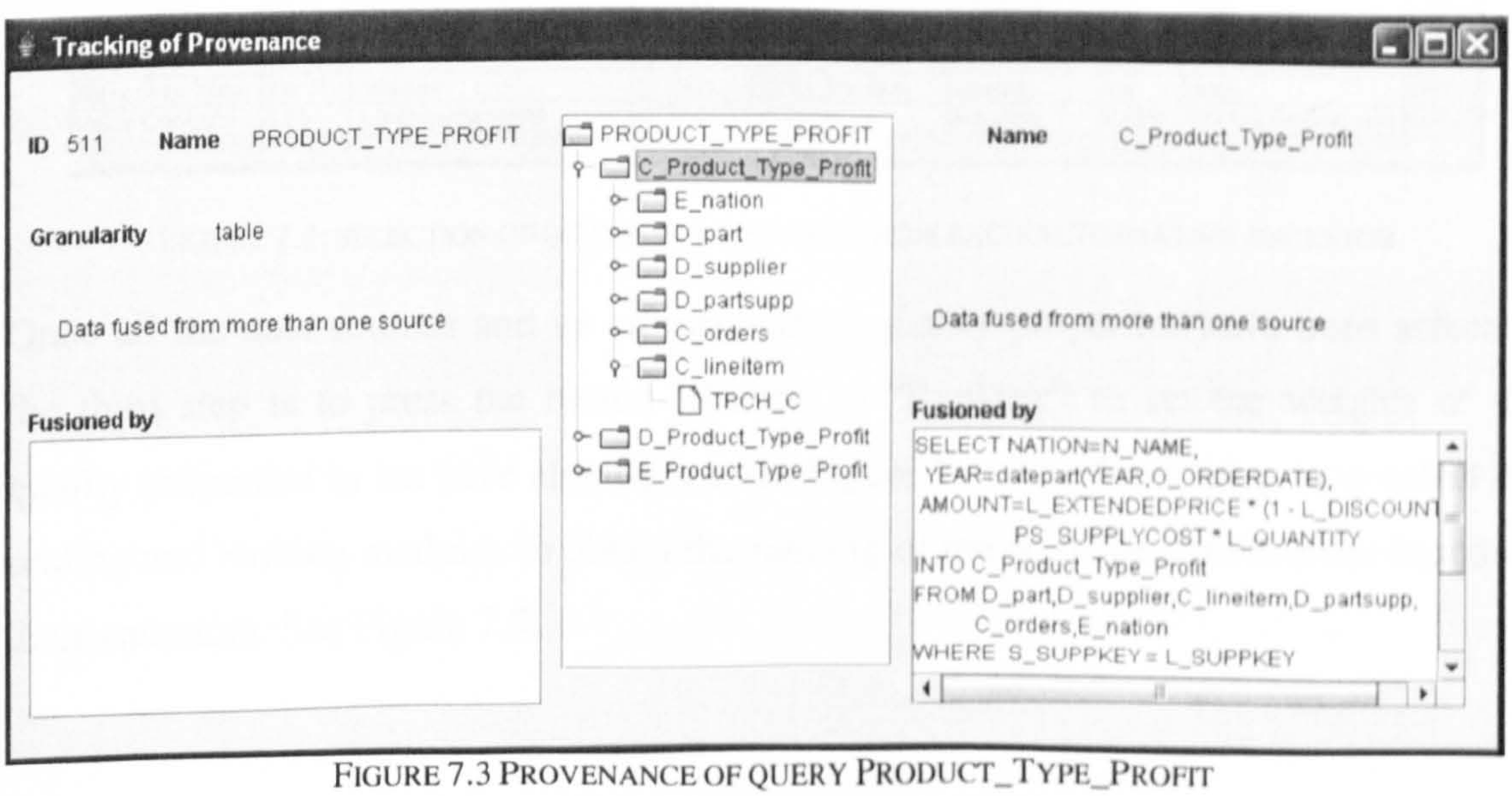


FIGURE 7.3 PROVENANCE OF QUERY PRODUCT_TYPE_PROFIT

Every time an ancestor is selected, all its available scores are displayed in the lower left table with a header "criterion score unit" showed in Figure 7.4. Consequently, the second step is to select any ancestors of interest and its desired quality properties by pressing the button labelled as ">>" to copy them to a lower right table with a header "Object Name Criterion Score Unit". As each query is executed on a number of data sources, there is a possibility to go through a lower level of granularity and test at attribute level and by provenance to obtain the relation which provides the attribute of

interest. For instance suppose AMOUNT as the attribute of interest and C_lineitem, D_lineitem and lineitem the corresponding ancestors from different data sources available. See figure 7.4.

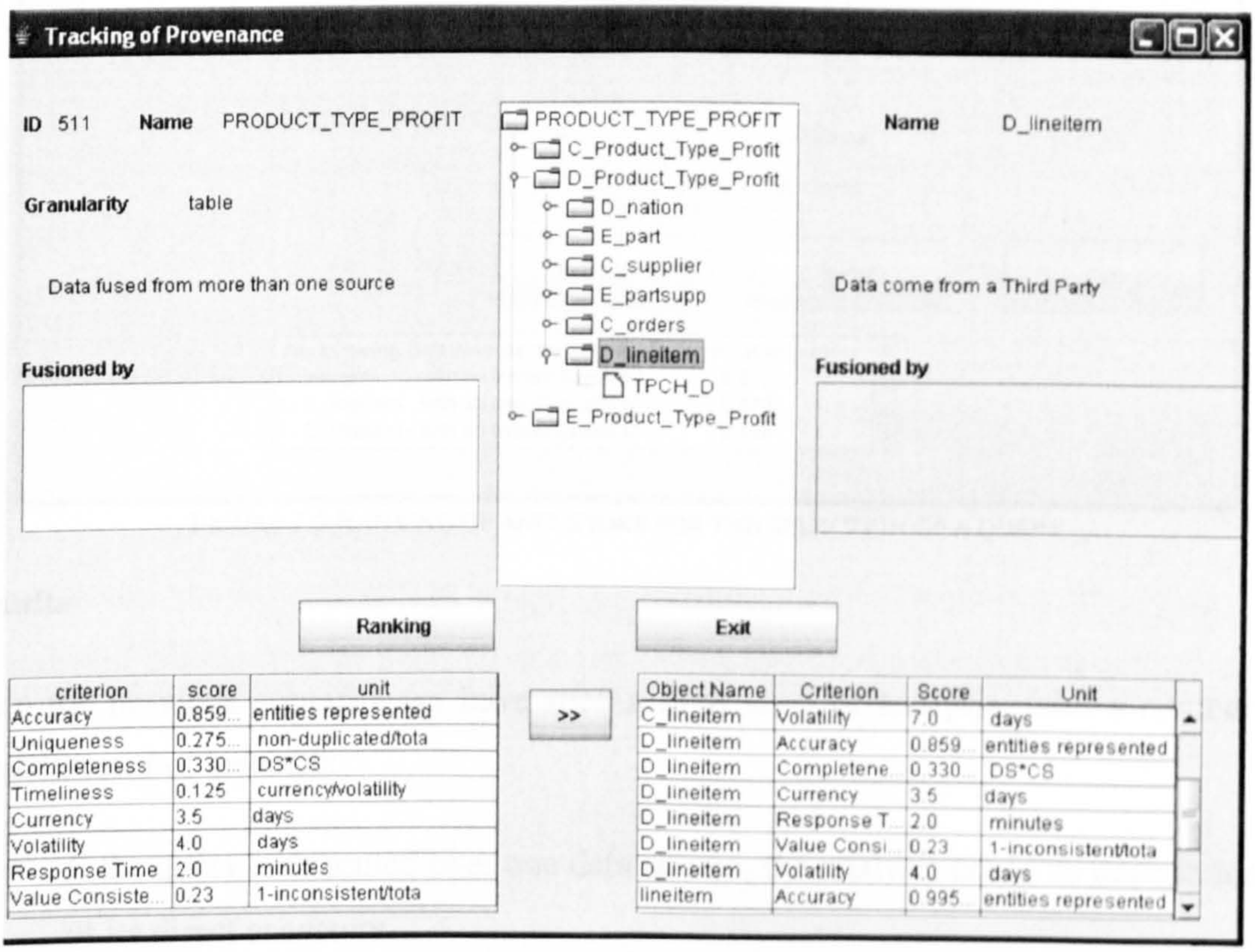


FIGURE 7.4: SELECTION OF QUALITY PROPERTIES FROM EACH ALTERNATIVE ANCESTOR

Once all the data sources and its corresponding quality properties have been selected, the third step is to press the button labelled as "Ranking" to set the weights of the quality properties as we have already seen in Figure 7.1. The fourth step is to select the scaling and ranking methods to obtain the ranking of the possible data sources based on their ancestors. See Figure 7.5.

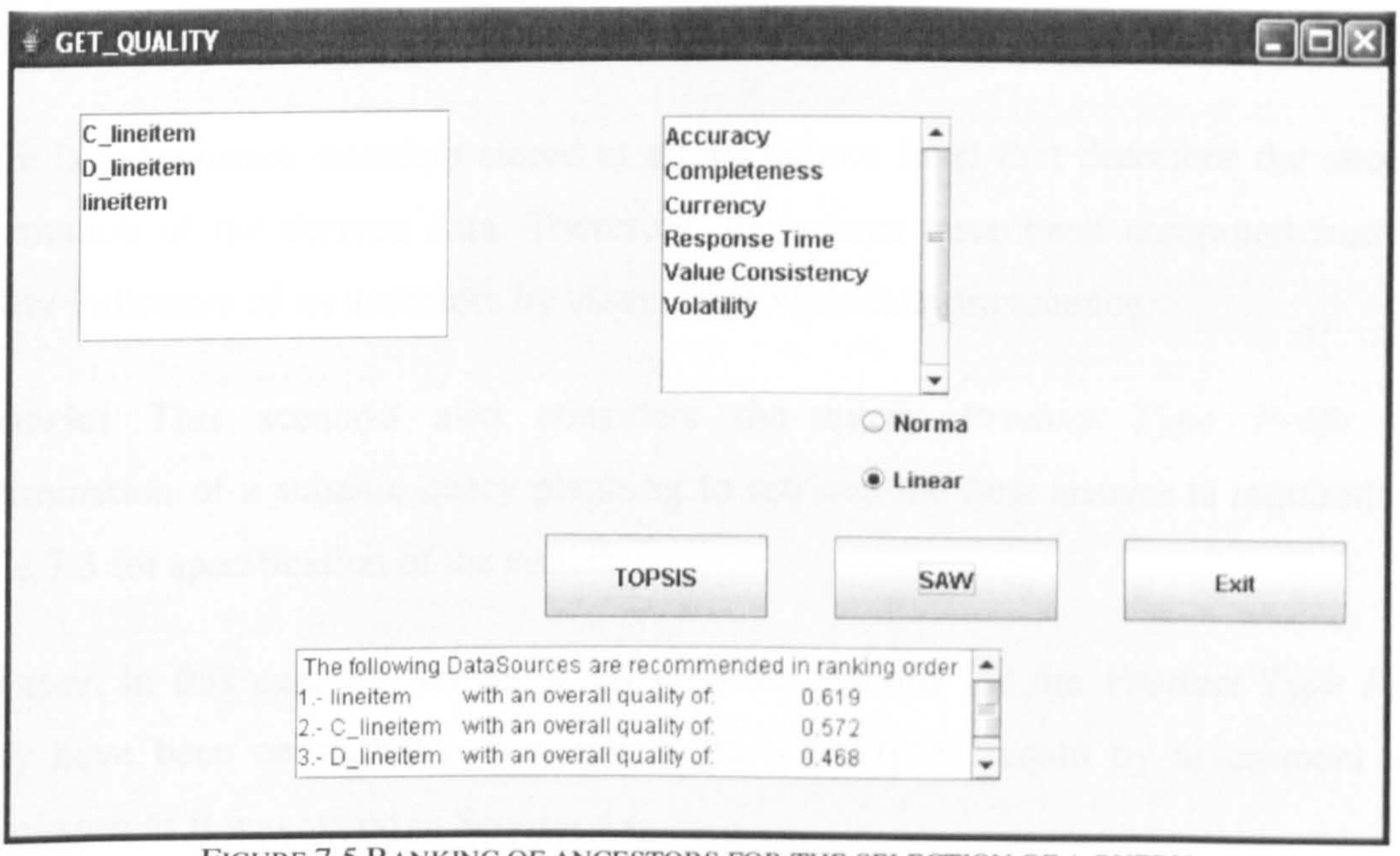


FIGURE 7.5 RANKING OF ANCESTORS FOR THE SELECTION OF A QUERY

Results

From the previous scenario, we have seen a high level of analysis, with a number of capabilities. For instance:

- If the query is executed over one data source, the analysis could be done in terms of its direct ancestors.

- If the query is executed over a number of data sources, we can select a specific ancestor relative to the attribute of interest.

- If the query contains attributes that are a result of a fusion of a number of attributes from different data sources, it is possible to go further and obtain the quality information of the elements that compound the attribute of interest.

In Test 1, the DQM recommends retrieving the AMOUNT attribute from the lineitem relation (see Fig. 7.5). If we compare Test 1 against Test 2, the lineitem relation was stored in a different database from TPCH_C (which was the suggested data source in Test 1). Considering the quality of ancestors, DQM recommends the execution of the query *Product Type Profit* from a different database. If we are interested in the correct estimation of AMOUNT, on the basis that it is a computed value from two attributes of the lineitem relation, then the latter is the best option. Therefore, we can conclude that the recommendation has changed according to the level of granularity assessed.

Test Case 3: Analysis of derived data based on its quality properties.

There is provenance metadata stored at a data source level that describes the ancestor information of the derived data. Therefore, the scores have been computed from the quality indicators of its ancestors by assessment with data provenance.

Scenario: This scenario also considers the query *Product Type Profit.* The determination of a suitable query planning to retrieve the best answer is required. See Table 7.3 for specification of the context.

However, in this case the scores of all possible options for the *Product Type Profit* query have been computed and stored in the Quality Metadata by assessment with provenance as it was stated in Section 4.6.

The first step within the DQM prototype is the selection of the application domain. The second step corresponds to the specification of the query. The third step corresponds to the identification of the quality properties and its priorities to obtain the ranking of queries based on their quality score.

As we have seen from previous scenarios, there is a possibility to type the identifier of the general name of the query in the federation and all the queries semantically equal will be presented (this feature is shown in scenario 2 in Figure 7.3).

Such alternatives are presented as `C_Product_Type_Profit`, `D_Product_Type_Profit` and, `E_Product_Type_Profit`.

At this point, a query is executed against the Provenance Metadata to obtain all the elements for the specified application domain. Users shall select which queries they need to evaluate along with the ranking and scaling methods. See Figure 7.1.

As can be observed in Figure 7.6, DQM suggests the execution of the query called `D_Product_Type_Profit`, for a better quality of data according with the user specifications.

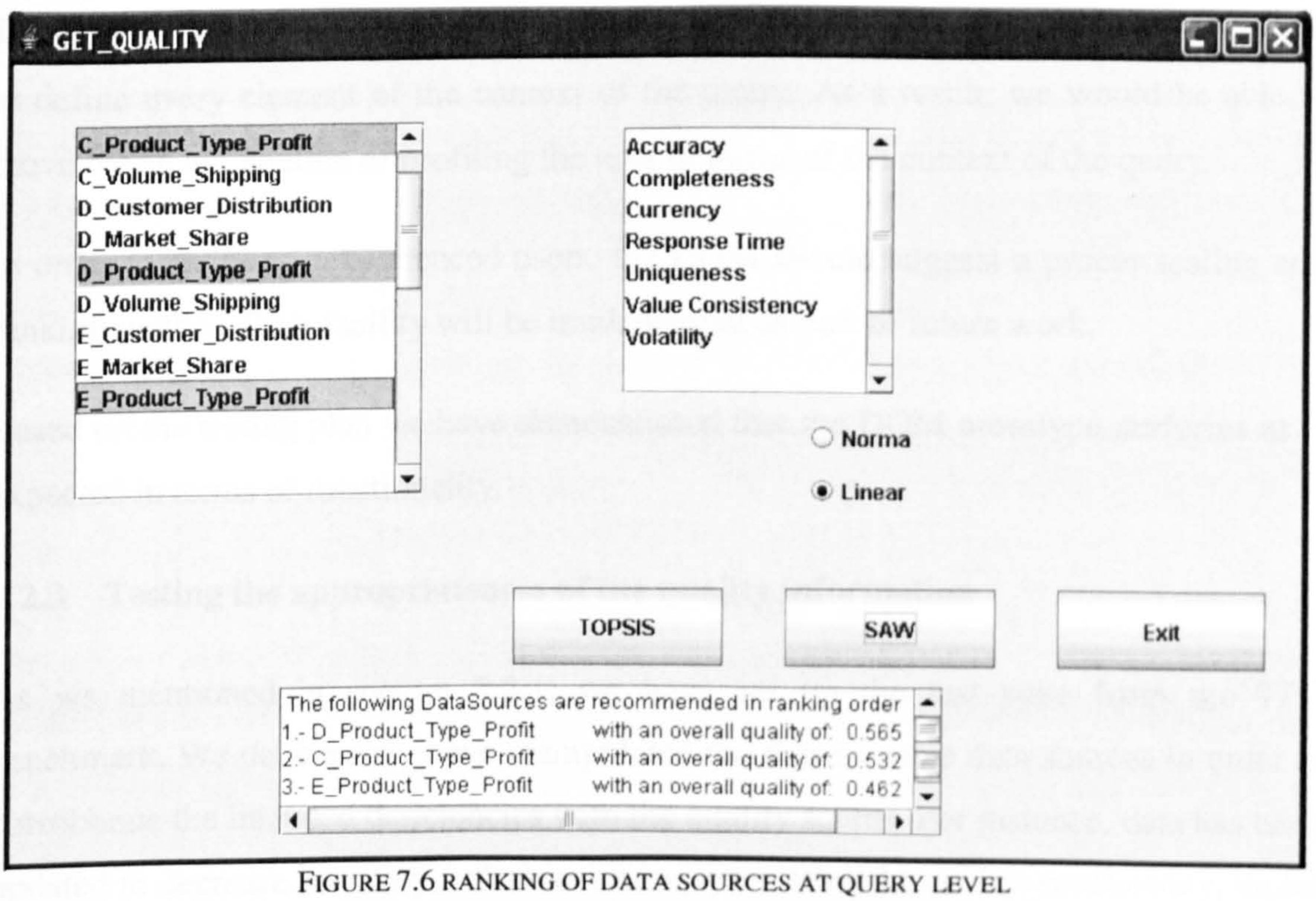


FIGURE 7.6 RANKING OF DATA SOURCES AT QUERY LEVEL

Results

Having quality metadata and provenance metadata available, the assessment of data sources can be performed for any data sources including integrated data, and ancestor's datasets at any granularity level.

Summary of outcomes

If we compare the three cases, we can conclude that quality estimation changes based on the granularity level under the same query event, as we can observe in Table 7.4.

Test Case	Quality Metadata	Provenance Metadata	Outcome	Granularity Level
1	√	X	TPCH_C	Database
2	Incomplete	Incomplete	lineitem	Relation
3	√	√	D_Product_Type_Profit	Query

TABLE 7.4 DQM OUTCOMES WITHIN THE TEST CASES.

The DQM deals with the conditions established by achieving the functionality of the three test cases.

The DQM establishes a mechanism by which we can provide support to users who want to define every element of the context of the query. As a result, we would be able to move for an automation of profiling the user in terms of the context of the query.

In order to support inexperienced users, the DQM should suggest a proper scaling and ranking method. This facility will be implemented as part of future work.

Based on the testing plan we have demonstrated that the DQM prototype performs as is expected in terms of functionality.

7.2.3 Testing the appropriateness of the quality information

As we mentioned in section 7.2.1, we have set up the test suite from the TPC benchmark. We deliberately have manipulated the data in some data sources in order to corroborate the intended deficiencies with the quality scores. For instance, data has been updated to decrease accuracy, references have been deleted to affect consistency, tuples have been duplicated to decrease uniqueness. We controlled the proportion of changes made on each source, to simulate known conditions which should generate expected responses, and then test if the system is providing those responses. By doing so, we could corroborate that the metrics were actually reflecting the quality of each data source.

Initial Conditions

The TPCH benchmark contains two relations called `customer` with a total of 750 tuples, and `orders` containing 7500 tuples. Appendix B.1.1 describes their corresponding schemas. Within the TPCH benchmark, there are programs that generate data automatically for loading data into the relations. We will assume that such values correspond to the real world. Therefore, the initial conditions are that there are no quality problems with an initial value of 1.0 per every data quality score. However, we have deliberately made some changes on both relations in order to reflect such changes in the quality measures. For comparison purposes, we have considered a range of $\pm 5\%$ of precision for the ranges estimated as suitable of been acepted within the expected range.

Applied changes to `customer`

The customer relation contains an attribute called C_CUSTKEY as its primary key. Such data values were generated automatically by a sequentially counter. Therefore, we can handle ranges of tuples by retrieving the even, odd, or multiples of any number from the primary key in order to control the number of changes, for the estimation of the expected scores. We have done so by using the `Modulo` operator which is part of the Transact-SQL extension. Refer to the Sybase TransactSQL Users' Guide for further detail [Transact02].The operator `Modulo` finds the integer remainder after a division involving two whole numbers. For example, 21% 11 = 10 because 21 divided by 11 equals 1 with a remainder of 10.

The weak relation accuracy is the number of tuples where every attribute is correct divided by the total number of rows (referred to in section 3.5.1). Therefore, we deliberately updated all rows, which belong to an even primary key (PK); that gives an estimation of half out of the total number of tuples considered as wrong when measuring accuracy. Consequently, the expected outcome should fall within the range of 45%-55% accurate tuples.

The weak relation completeness is the number of tuples with all its attributes filled with non-null values divided by the number of tuples (referred to in section 3.5.2). Therefore, inserting NULL values to any attribute to 40 tuples approximately, results in a 97% of complete tuples. Therefore, the result should be within a range of 92%-99% of complete tuples.

The consistency at the relation level is the percentage of tuples with all instances of the attributes consistent (referred to in section 3.5.3). Therefore, we deleted the foreign key value of the tuples which primary key value was a multiple of four, resulting in a range of 200 tuples updated, or 76% of consistent tuples. Accordingly, the expected range is within 71%-81% of consistent tuples.

As the percentage of unique tuples in a relation is the number of non-duplicated tuples divided by the cardinality of the relation, we duplicated the tuples with PK values were multiple of seven, which gives a range of a hundred duplicated tuples out of 750 results, or in other words 86% unique tuples. Hence, the acceptance range is within the range of 81%-91% of uniqueness.

Applied changes to `orders` and expected outcomes

The attribute O_ORDERKEY is the primary key of the orders relation. O_ORDERKEY contains values generated automatically by a counter within the TPCH benchmark through a C program. Therefore, in order to control the number of changes and to estimate the expected scores, we have manipulated its tuples by the Modulo operator.

For instance, in the case of the weak relation accuracy we deliberately updated all rows, which belong to an even PK; that gives an estimation of half out of the total number of tuples considered as wrong values. We also updated 15% of the odd values. Therefore, we expect between 30% to 40% of accurate tuples.

For the estimation of completeness, we inserted NULL values to any attribute value to approximately 576 out of 7500 tuples, which would result within a range of 87% to 97% of complete tuples.

In terms of consistency, we deleted the Foreign Key (FK) value of the tuples with even PK, resulting in approximately 3750 tuples updated, within a range of 45% to 55% inconsistent tuples.

In the case of uniqueness, we duplicated the tuples which PK value was multiple of seven, resulting to a thousand duplicated tuples out of 7500 results within a range of 81% to 91% of unique tuples.

The Table 7.5 shows the range of expected outcomes for accuracy, completeness, consistency, and uniqueness for customer and orders relations.

%	Accuracy	Completeness	Consistency	Uniqueness
Customer	45-55	92-99	71-81	81-91
Orders	30-40	87-97	45-55	81-91

TABLE 7.5 EXPECTED OUTCOMES

Computing the scores with the DQM prototype

In the DQM main menu, we require to select Measurement Model and click on the "Compute Scores" option. The type of Information System should be Decision Support (TPC-H benchmark in this case). Such measures will be stored in the Data Quality Metadata. See Figure 7.7 for the estimation of data quality scores by the DQM prototype.

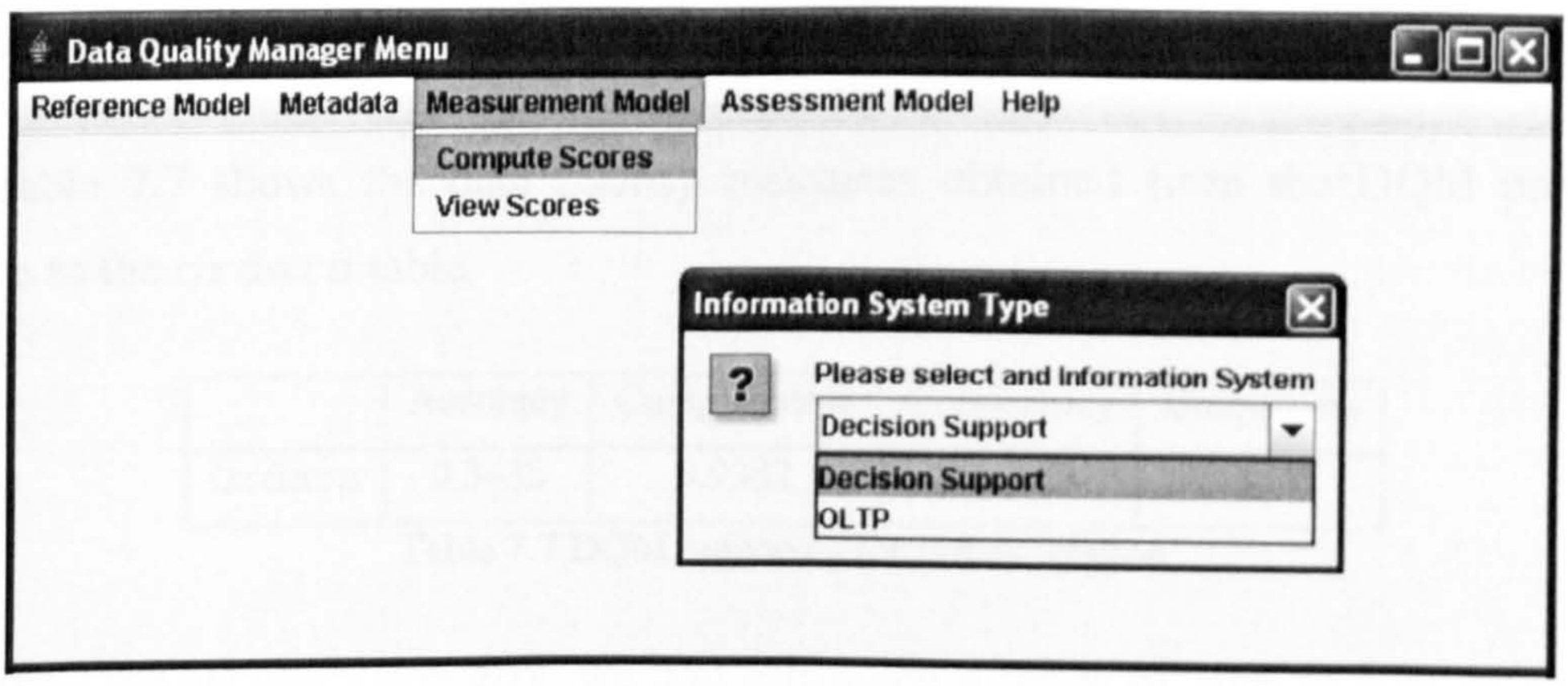


FIGURE 7.7 COMPUTING SCORES

Results

Outcomes for `Customer` relation

The Table 7.6 shows the data quality measures obtained from the DQM prototype relative to the `Customer` table.

	Accuracy	Completeness	Consistency	Uniqueness
`Customer`	0.4867	0.9618	0.76	0.8587

Table 7.6 Quality scores of table `Customer`

In the case of the accuracy property, we had estimated a range of 45% to 55% of accurate tuples, and the value produced by the prototype was 0.48, which was within the expected outcome.

For completeness, we estimated a value within the range of 92% to 99% of complete values and the prototype did produce a value of 0.96, which falled within the expected range.

As we can see, in the case of consistency we estimated a range of 71% to 81% of consistent tuples. The produced outcome of 0.76 falled within the expected range.

Regarding uniqueness, we estimated a range of 81% to 91% of unique values, the outcome falled within the range.

Outcomes for `orders` relation

The Table 7.7 shows the data quality measures obtained from the DQM prototype relative to the `orders` table.

	Accuracy	Completeness	Consistency	Uniqueness
`Orders`	0.3438	0.9382	0.485	0.8587

Table 7.7 DQM outcomes for orders relation

Regarding the accuracy property, we had estimated a range of 30% to 40% of accurate tuples, and the value produced by the prototype was 0.34, which is in the expected outcome.

In the case of completeness, we estimated a value within the range of 87% to 97% of complete tuples, and the prototype did produce a value of 0.93, which was expected.

For the consistency property, we expected a range of 45% to 55% of consistent tuples, such value is reflected with the produced outcome of 0.485.

In the case of uniqueness, we estimated a range of 81% to 91% of unique values, the outcome of 0.85 was within the range.

Conclusions:

What we are trying to demonstrate by running these experiments on these datasets is if the expected outcomes correspond with the actual values computed. From the results shown in Tables 7.6 and 7.7, we can conclude that the DQM prototype performs as expected in terms of the accuracy of the scores obtained. However, this is a limited indicative measurement of the appropriateness of the qualitative information of the DQM. Therefore, an extensive, rigorous testing is required as part of future work.

Summary of the outcomes

The results of these tests show the following:

- The prototype provides appropriate scores according with the expected outcomes based on the actual quality of data.

7.2.4 Testing the ranking of the data sources

We designed the set of tests proposed in this section to validate the ranking of the data sources given by the prototype according to their quality scores. As the scaling and ranking methods depend on the positive or negative nature of the quality criteria involved in the test, we have classified the tests as follows:

- Testing the ranking of data sources based on positive criteria only.
- Testing the ranking of data sources based on negative criteria.
- Testing the ranking of data sources based on positive and negative criteria.

The Table 7.8 shows three data sets and their corresponding scores for accuracy, uniqueness, consistency, currency, and response time.

Data Source	Accuracy (+)	Uniqueness (+)	Consistency (+)	Currency (-)	Response Time (-)
C	0.6549	0.7138	0.6431	6.93	1.75
D	0.7262	0.6435	0.6228	5.72	2.4
E	0.5392	0.6475	0.8412	2.72	1.5

Table 7.8 Data Quality scores for data sources C,D, and E.

In the case of positive criteria, the data source with best score will be the one with the greatest value. In the case of negative quality criteria, the data source with the lowest score value will be the best option. From the measures presented in Table 7.8, we can conclude which data source is the best option in terms of a specific data quality criterion. For instance, D is the most accurate data source. See Table 7.9, which shows the expected data sources ranking.

Quality Property	1st option	2nd option	3rd option
Accuracy	D	C	E
Uniqueness	C	E	D
Consistency	E	C	D
Currency	E	D	C
Response Time	E	C	D

Table 7.9 Expected ranking of data sources

Test Case 1 Testing the ranking of data sources based on positive criteria only.

Scenarios: An experienced user has three data sources (C, D, and E) available for query. The quality properties of interest are accuracy, uniqueness, and consistency. All the quality properties are positive.

We have considered three possible scenarios: 1) Accuracy is the highest priority; 2) Uniqueness is the highest priority; 3) Consistency is the highest priority.

The scaling and ranking methods utilised were Vector Normalization with SAW and Linear Scale Transformation with TOPSIS because the quality properties are all positive (referred to in Section 5.6).

The tests were executed in online and background modes with the following ranking outcomes shown in Table 7.10.

Quality Properties	Scenarios	Weights	Scaling/Ranking Methods	Ranking outcome 1st	2nd	3rd
+Accuracy	1	0.80	Vector N./SAW	D=0.63	C=0.59	E=0.50
		0.17 0.03	Linear/TOPSIS	D=0.92	C=0.62	E=0.038
+Uniqueness	2	0.17	Vector N./SAW	C=0.608	E=0.57	D=0.55
		0.80 0.03	Linear/TOPSIS	C=0.813	E=0.368	D=0.1
+Consistency	3	0.03	Vector N./SAW	E=0.657	C=0.541	D=0.52
		0.17 0.80	Linear/TOPSIS	E=0.929	C=0.117	D=0.034

Table 7.10 Ranking of data sources with positive criteria

Results

Scenario 1: The DQM suggests D as the most accurate data source with a SAW score of 0.63 and a TOPSIS score of 0.92, the next option is C with scores of 0.59 and 0.62 respectively.

Scenario 2: The DQM suggests C as the data source with less duplicated values with a SAW score of 0.608 and a TOPSIS score of 0.813, the next option is E with scores of 0.57 and 0.368 respectively.

Scenario 3: The DQM suggests E as the most consistent data source with a SAW score of 0.657 and a TOPSIS score of 0.929, the next option is C with scores of 0.541 and 0.117 respectively.

Comparing the expected outcomes on Table 7.9 against those shown on Table 7.10, we can conclude that the system behaves as expected for all the three conditions.

Test Case 2 Testing the ranking of data sources based on negative criteria.

Scenarios: An experienced user has three data sources (C, D, and E) available for query. The quality properties of interest are currency, and response time. All the quality properties are negative.

We have considered two possible scenarios: 1) Currency is the highest priority; 2) Response time is the highest priority.

The scaling and ranking methods utilised were Linear Scale Transformation with SAW and Vector Normalization with TOPSIS, because the quality properties are all negative (referred to in Section 5.6). The tests were executed in online and background modes with the following ranking outcomes shown in Table 7.11:

Quality Property	Scenarios	Weights	Scaling/Ranking Methods	Ranking outcome 1st	2nd	3rd
-Currency -Response Time	1	0.80	Vector /TOPSIS	E=1	D=0.278	C=0.098
		0.20	Linear/SAW	E=1	D=0.226	C=0.144
	2	0.20	Vector /TOPSIS	E=1	C=0.592	D=0.101
		0.80	Linear/SAW	E=1	C=0.578	D=0.056

Table 7.11 Ranking of data sources with negative criteria

Results

Scenario 1: The DQM suggests E as the most current data source with a TOPSIS score and SAW score of one, the next option is D with scores of 0.278 and 0.226 respectively.

Scenario 2: The DQM suggests E with the fastest response time with a TOPSIS score and SAW score of one. The next option is C with scores of 0.592 and 0.578 respectively.

Comparing the expected outcomes on Table 7.9 against those shown on Table 7.11, we can conclude that the system behaves as expected for all the three conditions.

Test Case 3 Testing the ranking of data sources based on negative and positive criteria.

An experienced user has three data sources (C, D, and E) available for query. The quality properties of interest are accuracy, and response time. Accuracy is a positive criterion, and response time is a negative criterion.

We have considered two possible scenarios: 1) Accuracy is the highest priority; 2) Response time is the highest priority.

The scaling and ranking methods utilised were Vector Normalization with TOPSIS and Linear Scale Transformation with SAW, because the quality properties are positive and negative (referred to in Section 5.6).

The tests were executed in online and background modes with the following ranking outcomes shown in Table 7.12:

Quality Property	Scenarios	Weights	Scaling/Ranking Methods	Ranking Outcome 1st	2nd	3rd
+Accuracy -Response Time	1	0.80	Vector /TOPSIS	D=0.723	C=0.632	E=0.277
		0.20	Linear/SAW	D=0.808	C=0.639	E=0.192
	2	0.20	Vector /TOPSIS	E=1	C=0.722	D=0
		0.80	Linear/SAW	E=1	C=0.722	D=0

Table 7.12 Ranking of data sources with positive and negative criteria

Results

Scenario 1: The DQM suggests D as the most accurate data source with a TOPSIS score of 0.723 and a SAW score of 0.808, the next option is C with scores of 0.632 and 0.639 respectively.

Scenario 2: The DQM suggests E as the fastest response time data source with a TOPSIS and SAW scores of one. The next option is C with a score of 0.722 with both methods.

Comparing the expected outcomes on Table 7.9 against those shown on Table 7.12, we can conclude that the system behaves as expected for all the three conditions.

Summary of the outcomes

The results of these tests show the following:

- The ranking of the data sources correspond to the expected ranking outcomes by changing the priority values of chosen quality criteria stated by the user.

Having obtained sensible expected outcomes, we can conclude that the prototype performs as is expected in terms of the accuracy of the answers it produces. However, these tests are at indicative level only. We have done the initial test only. We built a prototype, and did the experiments within a small range of tests where the DQM generates the information that we expected to. Therefore, the next stage will be to exhaustively test as part of future work.

7.3 Experimentation Plan

In order to test if the qualitative information produced by the DQM varies according to the context, we have designed a set of sanity checks in terms of quality criteria, quality priorities, levels of granularity and data provenance.

7.3.1 Experimental Hypotheses

The outcome that the DQM produces is that based on a set of criteria, and a query identified by the users, it provides qualitative information about the data sources involved in the query. Such qualitative information can be at a variety of levels of granularity, based on simple information from the data source itself; or based on the information stored in the metadata; or it can explicitly follow a provenance trail to find the original data sources. Hence, we identify the following experimental hypotheses:

1. The ranking of data sources changes according to the specification of quality criteria.
2. The ranking of data sources changes according to the specification of quality priorities.
3. The query outcome changes according to the assessment at different levels of granularity of their correspondent data sources or in other terms, the ranking of data sources changes according to the level of granularity assessed.
4. A query outcome varies with respect to either one of the following decision criteria: quality properties, assessment with data provenance, and no consideration of quality properties.

7.3.2 Assumptions

a) The sample of the n subjects or data sources has been randomly selected from the population it represents.

b) The original scores obtained for each of the subjects are in the format of interval/ratio data, see Section 3.5.

c) As a measurement of quality is directly related to the level of granularity, we conclude that scores measured at lower level of granularity will provide a greater degree of accuracy than aggregated scores produced at higher levels. See Section 3.5.9, for further detail.

d) Taking the best data source will result in the best query outcome for that specific context. Users are able to assign the context of the query in terms of the quality properties, quality priorities and the ranking and scaling methods according to their experience.

The first two assumptions are required for selecting the appropriate statistical procedure and samples according to Sheskin in [Sheskin04].

The second two assumptions correspond to the selection of the best data source according to the degree of accuracy required through the context specification. If the DQM provides qualitative information at multiple levels of granularity then it would be reasonable to utilise such information to deal with extensional inconsistencies. We therefore require to test if the DQM varies according to the context specification.

7.3.3 Types of Information System

The conducted experiments are based on two populations corresponding to the following benchmarks:

- The TPC Benchmark™H (TPC-H) already explained in section 7.2.2 and further discussed in Appendix B.
- The TPC Benchmark™C (TPC-C) already explained in section 7.2.2 and further discussed in Appendix B.

7.3.4 Sample Design

This Section identifies some basic concepts of sampling in order to explain how the samples were designed.

7.3.4.1 Dependent sample

A dependent sample design is also known as matched-subjects design [Gravetter00]. In this type of sampling, each of n subjects (data sources or queries) serves in each of the k experimental conditions. When a dependent sample design is conceptualized as a randomized-blocks design, it is because within each block the same subject is matched with itself by virtue of serving under the experimental conditions [Sheskin04].

7.3.4.2 Independent sample

In an independent sample design, each of the n different subjects is randomly assigned to one of the experimental group, with one group representing the experimental group and the other the control group.

As we are concerned with the ordering of the same set of data sources according to different conditions depending on the hypothesis to test, dependent sample design will be used for testing our hypothesis.

7.3.5 Variables

The independent variables are those, which cause changes or distinction between the groups of samples. Therefore, the identified independent variables are the Type of Information System, the quality properties, the quality priorities, the level of granularity, the ranking methods, the scaling methods, and the fusion function. The last independent variable has not been considered on the experimentation plan, because the assessment options and limitations already discussed in Section 4.6 relative to fused data. Such variable shall be included as part of future work.

The manipulation of the independent variables will represent different experimental conditions. The experiments will be controlled to rule out any confounding variable by changing one independent variable at a time.

As the conclusions of the experiments are related to the level of statistical significance change of the ranking of data sources or query, the dependent variable is the ranking of the data sources.

7.3.5.1 Quality dimensions

The quality dimensions used here were selected from the Data Quality Reference Model (See section 3.4) and correspond to Accuracy, Completeness, Consistency, Currency, Response Time, Timeliness, Uniqueness, and Volatility.

We have designed a number of conditions. Such conditions define which quality properties along with their priorities are representative depending on the hypothesis of the experiment. The quality properties will vary for hypothesis 1.

7.3.5.2 Weights

The quality priorities have been chosen according to each hypothesis. The quality priorities will vary for hypothesis 2.

7.3.5.3 Level of granularity

The assessment of data sources has been computed at database, table, attribute, and derived data or query as previously indicated in section 3.5, and section 4.7. The level of granularity is considered for hypothesis 3.

7.3.5.4 Ranking and Scaling Methods

The experiments will be conducted by using the following conventions:

- In the case that positive and negative criteria are involved in the decision matrix, we use the vector normalization scaling method with Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), or the scaling method linear scale transformation with Simple Additive Weighting (SAW).
- If there are only positive criteria in the decision matrix, we use vector normalization scaling method with Simple Additive Weighting (SAW).

This combination of scaling factors and ranking methods have been already discussed, and justified in Section 5.6.

7.3.6 Procedure for the Wilcoxon matched-pairs signed ranks test

As we are obtaining a rank-order data from the Multi-Attribute Decision Making methods then non-parametric procedures are the most suitable for our experiments [Sheskin04].

According to the characteristics of the samples obtained from the DQM, and the fact that the first three hypotheses relate to the ordering of data in two dependent populations, the experiments are conducted using the Wilcoxon matched-pairs signed-ranks test [Wilcoxon64].

In the case of hypothesis 4 concerned with the ordering of data in three dependent populations, the experiments are conducted using the Friedman's Test (Non-Parametric Repeated Measures Comparisons) which is discussed in Section 7.7.

7.3.6.1 Null versus Alternative Hypothesis

In order to evaluate our research hypotheses two statistical hypotheses are required: The null hypothesis, which is represented, by the notation *Ho*, and the alternative hypothesis, which is represented by the notation H_1.

The null hypothesis is a statement of no effect or no difference. In the underlying data sources represented by Condition 1 and Condition 2, the median of the difference scores (which will be represented by the notation θ_D) equals zero, no bound permitted.

Null hypothesis $Ho: \Theta_D = 0$

With respect to the sample data, this translates into the sum of the ranks of the positive difference scores being equal to the sum of the ranks of the negative difference scores i.e. $\sum T+ = \sum T-$.

The alternative hypothesis represents a statistical statement indicating the presence of an effect or a difference. In the underlying data sources represented by Condition 1 and Condition 2, the median of the difference scores is some value other than zero.

Alternative Hypothesis $H_1: \Theta_D \neq 0$

With respect to the sample data, this translates into the sum of the ranks of the positive difference scores not being equal to the sum of the ranks of the negative difference scores i.e. $\sum T+ \neq \sum T-$. This is a non-directional alternative hypothesis and it is evaluated with a two-tailed test. The reason of using a two-tailed test is that we are interested in the presence of any change in the order of the ranked data sources.

7.3.6.2 Test computations

A difference score is computed for each pair of matched subjects by subtracting a subject's score in Condition 2 from its score in Condition 1. The hypotheses evaluated with the Wilcoxon matched-pairs signed ranks test are whether or not the median of the difference scores equals zero. On the one hand, if a significant difference is obtained, it indicates there is a high likelihood the two samples/conditions represent two different population. In other words, the order of the data sources under two conditions is statistically significantly different and the alternative hypothesis is proven. On the other hand, if the median of the difference scores equals zero then the null hypothesis is true, and consequently there is no difference in the order of the data sources under two conditions. The calculus of the tests presented in this thesis have been executed using the statistics software called GraphPad InStat 3.06 Copyright © 1992-2003 by GraphPad Software Inc. which contains the statistical tables required. See reference [GraphPad] for further detail.

7.3.6.3 Interpretation of the test results

If the sample is derived from a population, in which the median of the difference scores equals zero, the values $\sum T+$ and $\sum T-$ will be equal to one another. If the value of $\sum T+$ is significantly greater than $\sum T-$, it indicates there is a high likelihood that Condition 1 represents a population with higher scores than the population represented in Condition 2. If the value of $\sum T-$ is significantly greater than $\sum T+$, it indicates there is a high likelihood that Condition 2 represents a population with higher scores than the population represented in Condition 1.

The question is however, whether the difference is significant i.e., whether it is large enough to conclude that is unlikely to be the result of a chance.

The absolute value of the smaller of the two values $\sum T+$ and $\sum T-$ is designated as the Wilcoxon T test statistic. The T value is interpreted by employing Table 2 (Table of Critical T values for Wilcoxon Signed Ranks and Matched-Pairs Signed Ranks Tests) in Appendix C. In order to be significant, the obtained value of T must be equal to or less than the tabled critical T value at the pre-specified level of significance.

Since the null hypothesis can only be rejected if the computed value T is equal to or less than the tabled critical value at the pre-specified level of significance, we can conclude that in order for the no directional alternative hypothesis $H_1 : \Theta_D \neq 0$ to be supported, it is irrelevant whether $\sum T+ > \sum T-$ or $\sum T+ < \sum T-$.

7.4 Set of Experiments for hypothesis 1

"The ranking of data sources changes according to the specification of quality criteria."

We have conducted a study to determinate whether or not the ranking of data sources changes according to the specification of quality criteria. The Data Quality Manager computes the ranking of each data source for each of the experimental conditions. Such conditions vary with respect to a different set of quality properties and same weights between them.

7.4.1 Experiment 1

Initial conditions

The independent variable in this experiment is the quality criteria, (shown Table 7.13).

Information System	Granularity Level	Ranking/Scaling Methods	Condition	Quality Criterion	Weight
TPCC	Table	TOPSIS	1	Accuracy Response Time	50 50
			2	Completeness Consistency	50 50

TABLE 7.13 EXPERIMENT CONDITIONS

Testing results

Wilcoxon matched-pairs signed-ranks test

Does the median of the differences between Acc-50/RT-50 and Com-50/VC50 differ significantly from zero?
The two-tailed p value is < 0.0001, considered extremely significant.

Calculation details

Sum of all signed ranks (W) = -306.00
Sum of positive ranks (T+) = 36.000
Sum of negative ranks (T-) = -342.00
Number of pairs = 27

Assumption test: Was the pairing effective?

Nonparametric Spearman correlation coefficient(r) = 0.3822
The one-tailed P value is 0.0246, considered significant.
Effective pairing results in a significant correlation between the columns. With these data, the pairing (or matching) appears to be effective.

Summary of Data

Parameter:	Acc-50/RT-50	Com-50/VC-50	Difference
Mean:	0.2667	0.5049	-0.2381
# of points:	27	27	27
Std deviation:	0.2261	0.2002	0.2710
Std error:	0.04351	0.03854	0.05215
Minimum:	0.000	0.0005460	-0.7175
Maximum:	0.7500	0.9376	0.4014
Median:	0.2563	0.5092	-0.2183
Lower 95% CI:	0.1773	0.4256	-0.3454
Upper 95% CI:	0.3562	0.5841	-0.1309

TABLE 7.14 TESTING RESULTS FOR HYPOTHESIS 1

Interpretation of the test results

Since the testing results show that $\sum T+ \neq \sum T-$ we can conclude that there is a difference in the ranking of the data sources according to the specification of different quality criteria. However, the null hypothesis can only be rejected if the computed T = 36 is equal or smaller than the tabled critical value at the level of significance p = 0.0001. From the testing results, we can conclude that there is an extremely significant probability that the ranking of the data sources is different according with the quality criteria. Moreover, the effectiveness of the pairing of the data sources was estimated by the Spearman correlation test with a significance of 0.02 from Table 3 of Critical Values for Spearman's Rho (r) in Appendix C.

7.4.2 Experiment 2

Initial conditions

The experiment conditions correspond to the use of either accuracy or timeliness as the only one quality criterion for the ranking of data sources as can be seen in Table 7.15.

Information System	Granularity Level	Ranking/Scaling Methods	Condition	Quality Criterion	Weight
TPCC	Table	TOPSIS	1	Accuracy	100
			2	Timeliness	100

TABLE 7.15 EXPERIMENT CONDITIONS FOR HYPOTHESIS 1

Testing results

Wilcoxon matched-pairs signed-ranks test

Does the median of the differences between Accuracy-100 and Timeliness-100 differ significantly from zero?
The two-tailed P value is < 0.0001, considered extremely significant.
Calculation details
Sum of all signed ranks (W) = -322.00
Sum of positive ranks (T+) = 28.000
Sum of negative ranks (T-) = -350.00
Number of pairs = 27
Assumption test: Was the pairing effective?
Nonparametric Spearman correlation coefficient (r) = 0.1539
The one-tailed P value is 0.056, considered significant. The correlation coefficient indicates that the pairing or matching appears to be effective.

Summary of Data

Parameter:	Accuracy-100	Timeliness-100	Difference
Mean:	0.05193	0.4517	-0.3997
# of points:	27	27	27
Std deviation:	0.1909	0.4151	0.4944
Std error:	0.03674	0.07989	0.09515
minimum:	0.000	0.000	-1.000
Maximum:	1.000	1.000	0.9545
Median:	0.01543	0.3481	-0.3305
Lower 95% CI:	-0.02362	0.2874	-0.5953
Upper 95% CI:	0.1275	0.6159	-0.2041

TABLE 7.16 TESTING RESULTS FOR HYPOTHESIS 1

Interpretation of the test results

Since the testing results shown that $\sum T+ \neq \sum T-$ we can conclude that there is a difference in the ranking of the data sources according to the specification of different quality criteria. Then, the null hypothesis can only be rejected if the computed T = 28 is equal to or smaller than the tabled critical value which is true at the level of significance p = 0.0001. Consequently, we can conclude that there is an extremely significant probability that the ranking of the data sources is different with respect to the quality properties considered.

7.4.3 Experiment 3

Initial conditions

The experiment conditions correspond to the use of either response time or accuracy as the only one quality criterion for the ranking of data sources (see Table 7.17).

Information System	Granularity Level	Ranking/Scaling Methods	Condition	Quality Criterion	Weight
TPCH	Table	TOPSIS	1	Response Time	100
			2	Accuracy	100

TABLE 7.17 EXPERIMENT CONDITIONS FOR HYPOTHESIS 1

Testing results

Wilcoxon matched-pairs signed-ranks test

Does the median of the differences between ResponseTime100 and Accuracy100 differ significantly from zero?

The two-tailed P value is < 0.0001, considered extremely significant.
(The P value is an estimate based on a normal approximation.)

Calculation details
Sum of all signed ranks (W) = 516.00
Sum of positive ranks (T+) = 522.00
Sum of negative ranks (T-) = -6.000
Number of pairs = 32

Assumption test: Was the pairing effective?
Nonparametric Spearman correlation coefficient (r)= 0.1506
The one-tailed P value is 0.053, considered significant.
Effective pairing results in a significant correlation between the columns.
With these data, the pairing (or matching) appears to be effective.

Summary of Data

Parameter:	ResponseTime100	Accuracy100	Difference
Mean:	0.8509	0.1086	0.7423
of points:	32	32	32
Std deviation:	0.1899	0.1417	0.2781
Std error:	0.03358	0.02506	0.04916
Minimum:	0.000	5.660E-05	-0.5255
Maximum:	1.000	0.5255	0.9285
Median:	0.8929	0.08118	0.8339
Lower 9.5% CI:	0.7824	0.05745	0.6421
Upper 95% CI:	0.9194	0.1597	0.8426

TABLE 7.18 TESTING RESULTS FOR HYPOTHESIS 1

Interpretation of the test results

The testing results show that $\sum T+ = 522$ and , so we can conclude that there is a difference in the ranking of the data sources according to the specification of quality criteria. Besides, the null hypothesis can only be rejected if the computed T = 6 is equal or smaller than the tabled critical value at the level of significance p = 0.0001. We can conclude that there is an extremely significant probability that the ranking in the data sources is different from different quality criteria specifications.

7.4.4 Conclusions

As all experiments have shown highly statistically significant changes in the ranking according to different quality criteria with fixed values of other possible factors that could affect our results, we accept the hypothesis 1.

"The ranking of data sources changes according to the specification of quality criteria."

7.5 Set of Experiments for Hypothesis 2

"The ranking of data sources changes according to the specification of quality priorities."

We have conducted a study to determinate whether or not the ranking of data sources changes according to the specification of quality priorities. Consequently, the experimental conditions in this set of experiments are based on using the same set of quality properties but changing their corresponding weights.

7.5.1 Experiment 1

Initial conditions

The experiment conditions are shown in Table 7.19.

Information System	Granularity Level	Ranking Method	Quality Criteria	Weights for Condition 1	Weights for Condition 2
TPCC	Table	TOPSIS	Accuracy	100	100
			Completeness	65	67
			Currency	40	78
			Response Time	30	43
			Uniqueness	90	95
			Consistency	80	90
			Volatility	50	80

TABLE 7.19 EXPERIMENT CONDITIONS FOR HYPOTHESIS 2

Testing results

Wilcoxon matched-pairs signed-ranks test

Does the median of the differences between BD1TOP and BD2TOP differ significantly from zero?

The two-tailed P value is < 0.0001, considered extremely significant

Calculation details

Sum of all signed ranks (W) = -356.00
Sum of positive ranks (T+) = 11.000
Sum of negative ranks (T-) = -367.00
Number of pairs = 27

Assumption test: Was the pairing effective?

Nonparametric Spearman correlation coefficient(r)= 0.8233
The one-tailed P value is < 0.0001, considered extremely significant
Effective pairing results in a significant correlation between the columns.
With these data, the pairing (or matching) appears to be effective.

Summary of Data

Parameter:	BD1TOP	BD2TOP	Difference
Mean:	0.1650	0.1929	-0.02798
# of points:	27	27	27
Std deviation:	0.1426	0.1338	0.01951
Std error:	0.02745	0.02575	0.003755
Minimum:	0.03915	0.05346	0.05447
Maximum:	0.6004	0.5724	0.02803
Median:	0.08480	0.1270	-0.02961
Lower 95% CI:	0.1085	0.1400	-0.03570
Upper 95% CI:	0.2214	0.2459	-0.02026

TABLE 7.20 TESTING RESULTS FOR HYPOTHESIS 2

Interpretation of the test results

Since the testing results show that $\sum T+=11$ is different from $\sum T-=-367$ we can conclude that there is a difference in the ranking of the data sources according to the specification of different quality priorities, even for the same set of quality properties. As the null hypothesis can be rejected due the computed T = 11 is smaller than the tabled critical value at 0.0001 level of significance, we can conclude that there is an extremely significant probability that the ranking in the data sources varies due to different quality priority specifications.

7.5.2 Experiment 2

Initial conditions

Information System	Granularity Level	Ranking Method	Quality Criteria	Weights for Condition 1	Weights for Condition 2
TPCC	Table	TOPSIS	Accuracy	100	50
			Completeness	65	50
			Currency	40	50
			Response Time	30	50
			Uniqueness	90	50
			Consistency	80	50
			Volatility	50	50

TABLE 7.21 EXPERIMENT CONDITIONS FOR HYPOTHESIS 2

Testing results

Wilcoxon matched-pairs signed-ranks test

Does the median of the differences between BD1TOP and BD2TOP differ significantly from zero?

The two-tailed P value is < 0.0001, considered extremely significant

Calculation details
Sum of all signed ranks (W) = -344.00
Sum of positive ranks (T+) = 17.000
Sum of negative ranks (T-) = -361.00
Number of pairs = 27

Assumption test: Was the pairing effective?
Nonparametric Spearman correlation coefficient (r) = 0.9126
The one-tailed P value is < 0.0001, considered extremely significant
Effective pairing results in a significant correlation between the columns.
With these data, the pairing (or matching) appears to be effective.

Summary of Data

Parameter:	BD1TOP	BD2TOP	Difference
Mean:	0.1650	0.2318	-0.06684
# of points:	27	27	27
Std deviation:	0.1426	0.1262	0.04368
Std error:	0.02745	0.02430	0.008407
Minimum:	0.03915	0.1036	-0.1159
Maximum:	0.6004	0.5200	0.08048
Median:	0.08480	0.1940	-0.07083
Lower 95% CI:	0.1085	0.1818	-0.08413
Upper 95% CI:	0.2214	0.2817	-0.04956

TABLE 7.22 TESTING RESULTS FOR HYPOTHESIS 2

Interpretation of the test results

The testing results show that $\sum T+=17$ is different from $\sum T-=-361$ we can conclude that there is a difference in the ranking of the data sources according to the specification of different quality priorities, even for the same set of quality properties. As the null hypothesis can be rejected due to the absolute smaller value of T correspond to 0.0001 level of significance, we can conclude that there is an extremely significant probability that the ranking in the data sources varies due to different quality priority specifications.

7.5.3 Experiment 3

Initial conditions

Information System	Granularity Level	Ranking Method	Quality Criteria	Weights for Condition 1	Weights for Condition 2
TPCC	Table	SAW	Accuracy	100	50
			Completeness	65	50
			Currency	40	50
			Response Time	30	50
			Uniqueness	90	50
			Consistency	80	50
			Volatility	50	50

TABLE 7.23 EXPERIMENT CONDITIONS FOR HYPOTHESIS 2

Testing results

Wilcoxon matched-pairs signed-ranks test

Does the median of the differences between BD1SAW and BD2SAW differ significantly from zero?
The two-tailed P value is < 0.0001, considered extremely significant

Calculation details
Sum of all signed ranks (W) = -352.00
Sum of positive ranks (T+) = 13.000
Sum of negative ranks (T-) = -365.00
Number of pairs = 27

Assumption test: Was the pairing effective?
Nonparametric Spearman correlation coefficient (r)= 0.9664
The one-tailed P value is < 0.0001, considered extremely significant Effective pairing results in a significant correlation between the columns. With these data, the pairing (or matching) appears To be effective.

Summary of Data

Parameter:	BD1SAW	BD2SAW	Difference
Mean:	0.2174	0.2649	-0.04753
# of points:	27	27	27
Std deviation:	0.2160	0.2220	0.03592
Std error:	0.04157	0.04272	0.006914
Minimum:	0.05288	-0.09582	0.09218
Maximum:	0.7517	0.8309	0.03406
Median:	0.1172	0.1971	0.04147
Lower 95% CI:	0.1319	0.1771	-0.06175
Upper 95% CI:	0.3028	0.3528	-0.03332

TABLE 7.24 TESTING RESULTS FOR HYPOTHESIS 2

Interpretation of the test results

The testing results show that $\sum T+ = 13$ is different from $\sum T- = -365$, so we can conclude that there is a difference in the ranking of the data sources according to the specification of different quality priorities, even for the same set of quality properties. As the null hypothesis can be rejected due to the absolute smaller value of T correspond to 0.0001 level of significance, we can conclude that there is an extremely significant probability that the ranking in the data sources varies due to different quality priority specifications.

7.5.4 Conclusions

As all the experiments have shown highly statistically significant changes in the ranking according to different quality priorities using the same quality criteria properties using fixed values of other possible factors that could affect our results, we accept the hypothesis 2.

"The ranking of data sources changes according to the specification of quality priorities."

7.6 Set of experiments for Hypothesis 3

"Query outcomes change according to the assessment at different levels of granularity of their correspondent data sources."

The same query can be obtained from different data sources. The most common procedure to decide where queries should be executed from is by trusting one source of data against other. In the best case this trustiness is based on quality measures. However, in the case of complex queries which are executed against to a number of data sources, the quality of queries shall be determined by the assessment of their corresponding ancestors at a finer level of granularity for a better approximation of data quality.

We have conducted the following set of experiments to determinate whether or not the query outcomes change according to the assessment at different levels of granularity, at database level with no provenance against query level considering. Consequently, the experimental conditions in this set of experiments are based on using quality measures obtained for the query against the measures obtained at data base level.

7.6.1 Experiment 1

Initial conditions

Information System	Quality Criteria	Weights	Ranking/Scaling Methods	Condition	Granularity Level
TPCH	Accuracy	35	TOPSIS	1	Query
	Completeness	55		2	Database
	Timeliness	95			
	Uniqueness	25			
	Consistency	75			

TABLE 7.25 EXPERIMENT CONDITIONS FOR HYPOTHESIS 3

Testing results

Ranking of Data Sources with provenance against no provenance
Wilcoxon matched-pairs signed-ranks test

Do the median of the differences between Condition1-Prov and Condition2-NoPr differ significantly from zero?
The two-tailed P value is 0.0002, considered extremely significant.

Calculation details
Sum of all signed ranks (W) = -116.00
Sum of positive ranks (T+) = 2.000
Sum of negative ranks (T-) = -118.00
Number of pairs = 15

Assumption test: Was the pairing effective?
Nonparametric Spearman correlation coefficient (r) = 0.7893
The two-tailed P value is 0.0002, considered extremely significant.
Effective pairing results in a significant correlation between the columns. With these data, the pairing (or matching) appears to be effective.

Summary of Data

Parameter:	Condition1-Prov	Condition2-NoPr	Difference
Mean:	0.2597	0.3920	-0.1323
# of points:	15	15	15
Std deviation:	0.2510	0.2517	0.1597
Std error:	0.06480	0.06499	0.04125
Minimum:	0.09734	0.1742	-0.5872
Maximum:	0.9187	0.9225	0.01142
Median:	0.1756	0.2828	-0.08583
Lower 95% CI:	0.1207	0.2526	-0.2207
Upper 95% CI:	0.3987	0.5313	-0.04380

TABLE 7.26 TESTING RESULTS FOR HYPOTHESIS 3

Interpretation of the test results

Since the testing results show that $\sum T+ \neq \sum T-$ we can conclude that there is a difference in the ranking of the data sources considering provenance against data sources only. Furthermore the null hypothesis can be rejected due to the computed T = 2 is smaller than the tabled critical value at the level of significance p = 0.0002. From the testing results we can conclude then, that there is an extremely significant probability that ranking data sources considering provenance is different from ranking data sources with no consideration of its ancestors. Moreover, the effectiveness of the pairing of the data sources was estimated by the Spearman correlation test with a significance of 0.0002.

7.6.2 Experiment 2

Initial conditions

Information System	Quality Criteria	Weights	Ranking Method	Condition	Granularity Level
TPCH	Accuracy	75	TOPSIS	1	Query
	Completeness	25		2	Database
	Timeliness	30			
	Uniqueness	80			
	Consistency	35			

TABLE 7.27 EXPERIMENT CONDITIONS FOR HYPOTHESIS 3

Testing results

Ranking of Data Sources with provenance against no provenance
Wilcoxon matched-pairs signed-ranks test

Does the median of the differences between Condition1-Prov and Condition2-NoPr differ significantly from zero?

The two-tailed P value is 0.0004, considered extremely significant.

Calculation details
Sum_all signed ranks (W) = -112.00
Sum positive ranks (T+) = 4.000
Sum of negative ranks (T-) = -116.00
Number of pairs = 15

Assumption test: Was the pairing effective?
Nonparametric Spearman correlation coefficient (r) = 0.8464
The two-tailed P value is < 0.0001, considered extremely significant.
Effective pairing results in a significant correlation between the columns. With these data, the pairing (or matching) appears effective.

Summary of Data

Parameter:	Condition1-Prov	Condition2-NoPr	Difference
Mean:	0.4163	0.5666	-0.1503
Number of points:	15	15	15
Std deviation:	0.1728	0.1850	0.1083
Std error:	0.04461	0.04776	0.02798
Minimum:	0.1465	0.3170	-0.3689
Maximum:	0.7952	0.8435	0.07503
Median:	0.4408	0.5838	-0.1623
Lower 95% CI:	0.3206	0.4642	-0.2103
Upper 95% CI:	0.5120	0.6690	-0.09027

TABLE 7.28 TESTING RESULTS FOR HYPOTHESIS 3

Interpretation of the test results

The testing results show that $\sum T+ = 4$ is different from $\sum T- = -116$ with a level of significance p = 0.0004 we can conclude that there is a difference in the ranking of the data sources considering provenance against considering data as prime data source only. Moreover, the effectiveness of the pairing of the data sources was estimated by the Spearman correlation test with a significance of 0.0001.

7.6.3 Experiment 3

Initial conditions

Information System	Quality Criteria	Weights	Ranking Method	Condition	Granularity Level
TPCH	Accuracy	95	TOPSIS	1	Query
	Completeness	75		2	Database
	Timeliness	55			
	Uniqueness	35			
	Consistency	15			

TABLE 7.29 EXPERIMENT CONDITIONS FOR HYPOTHESIS 3

Testing results

Ranking of Data Sources with provenance against no provenance
Wilcoxon matched-pairs signed-ranks test

Do the median of the differences between Condition1-Prov and Condition2-NoPr differ significantly from zero?

The two-tailed P value is 0.0002, considered extremely significant.

Calculation details

Sum of all signed ranks (W) = -116.00
Sum of positive ranks (T+) = 2.000
Sum of negative ranks (T-) = -118.00
Number of pairs = 15

Assumption test: Was the pairing effective?

Nonparametric Spearman correlation coefficient (r) = 0.7750
The one-tailed P value is 0.0003, considered extremely significant. Effective pairing results in a significant correlation between the columns With these data, the pairing (or matching) appears to be effective.

Summary of Data

Parameter:	Condition1-Prov	Condition2-NoPr	Difference
Mean:	0.2990	0.4461	-0.1471
# of points:	15	15	15
Std deviation:	0.2313	0.2175	0.1411
Std error:	0.05973	0.05615	0.03643
Minimum:	0.09457	0.2157	-0.5469
Maximum:	0.9208	0.9249	0.01050
Median:	0.2286	0.3532	-0.1146
Lower 95% CI:	0.1709	0.3257	-0.2253
Upper 95% CI:	0.4271	0.5666	-0.06901

TABLE 7.30 TESTING RESULTS FOR HYPOTHESIS 3

Interpretation of the test results

The testing results show that $\sum T+ \neq \sum T-$. Hence, we can conclude that there is a difference in the ranking of the data sources considering provenance against data sources only. Furthermore the null hypothesis can be rejected due to the computed T = 2 is smaller than the tabled critical value at the level of significance p = 0.0002. From the testing results we can conclude then, that there is an extremely significant probability that ranking data sources considering provenance is different from ranking data sources with no consideration of its ancestors. Moreover, the effectiveness of the pairing of the data sources was estimated by the Spearman correlation test with a significance of 0.0003.

7.6.4 Conclusions

All the experiments have shown highly statistically significant changes in the ranking of queries when the assessment has been performed at a finer level of granularity by data provenance against those data sources assessed with no consideration of data provenance. Consequently, we have enough evidence to accept the hypothesis 3.

"Query outcomes change according to the assessment at different levels of granularity of their correspondent data sources."

7.7 Procedure for the Friedman's Test

In order to test hypothesis 4, a comparison of three samples is required, the ranking of the query, the ranking of the database, and the sample corresponding to the set of data sources with no quality differences among them. In this case, there is a test for nonparametric repeated measures comparisons named the Friedman Test.

This test like Wilcoxon's test uses the ranks of the data to calculate the statistic.

7.7.1 Null versus Alternative Hypothesis

The notation *Ho* represents the null hypothesis and it is a statement of no effect or difference. In other words, the distributions are the same across repeated measures.

$$\textbf{Null Hypothesis } Ho: \Theta_{D1} = \Theta_{D2} = \Theta_{D3}$$

On the other hand, the notation H_1 represents the alternative hypothesis, and it represents a statistical statement indicating the presence of an effect or difference. In other words, the distributions across repeated measures are different, or at least one is not equal.

$$\textbf{Alternative Hypothesis } H_1 : Not\ H_0$$

These hypotheses are also expressed as comparing mean ranks across measures. The test statistic for the Friedman's test is a Chi-square with k-1 degrees of freedom (df), where k is the number of repeated measures. The Table of the Chi-Square distribution is shown in Appendix C Section 4. When the p-value for this test is small (usually <0.05) there is evidence to reject the null hypothesis.

When the p-value is low, there is evidence to reject *Ho*, and conclude that there is a difference between mean ranks.

7.8 Set of experiments for Hypothesis 4

"A Query outcome varies with respect to either one of the following decision criteria: quality properties, assessment with data provenance, and no consideration of quality properties."

The first condition is related to samples, which scores are estimated from the assessment with provenance, the second condition corresponds to scores calculated as if data was original, and the third condition is concerned with samples with no quality scores, having then the same score for every data source.

7.8.1 Experiment 1

Initial conditions

Information System	Quality Criteria	Weights	Ranking Method	Condition	Granularity Level
TPCH	Accuracy	35	TOPSIS	1	Query
	Completeness	55		2	Database
	Timeliness	95			
	Uniqueness	25			
	Consistency	75			
	None	None		3	Database

TABLE 7.31 EXPERIMENT CONDITIONS FOR HYPOTHESIS 4

Testing results

Ranking of Data Sources Provenance-NoProvenance-NoQuality

Friedman Test (Nonparametric Repeated Measures ANOVA)

The P value is < 0.0001, considered extremely significant.
Variation among column medians is significantly greater than expected by chance.
The P value is approximate (from chi-square distribution) because exact calculations would have taken too long.

Calculation detail

Group	Sum of Ranks
Provenance	30.000
NoProvenance	44.000
NoQuality	16.000

Number of Rows = 15
Number of Columns = 3
Friedman Statistic Fr = 26.133

Dunn's Multiple Comparisons Test

If the difference between rank sum means is greater than 13.117 then the p value is less than 0.05.

Comparison	Rank Sum Difference		Pvalue
Provenance vs. NoProvenance	-14.000	*	P<0.05
Provenance vs. NoQuality	14.000	*	P<0.05
NoProvenance vs. NoQuality	28.000	***	p<0.001

TABLE 7.32 TESTING RESULTS FOR HYPOTHESIS 4

Interpretation of testing results

This analysis indicates that there is a difference in the ranking across the use of quality properties, assessment considering provenance and no consideration of any quality property with $p < 0.0001$. For instance, the multiple comparisons indicate (at the 0.05 significance level) that the scores of the data sources with no consideration of provenance were greater than scores of data sources which were assessed by their ancestors. Note that the p-value for the Chi-square test is reported even though the sample size is small. In this case, the tabled value agrees with the Chi-square value. * = significant, ** = highly significant, *** = extremely significant, ns = no significant.

7.8.2 Experiment 2

Initial conditions

Information System	Quality Criteria	Weights	Ranking Method	Condition	Granularity Level
TPCH	Accuracy	75	TOPSIS	1	Query
	Completeness	25		2	Database
	Timeliness	30			
	Uniqueness	80			
	Consistency	35			
	None	None		3	Database

TABLE 7.33 EXPERIMENT CONDITIONS FOR HYPOTHESIS 4

Testing results

Ranking of Data Sources Provenance- NoProvenance-NoQuality

Friedman Test (Nonparametric Repeated Measures ANOVA)

The P value is < 0.0001, considered extremely significant.
Variation among column medians is significantly greater than expected by chance. The P value is approximate (from chi-square distribution) because exact calculations would have taken too lang.

Calculation detail

Group	Sum of Ranks
Provenance	17.000
NoProvenance	33.000
NoQuality	40.000

Number of Rows = 15
Number of Columns = 3
Friedman Statistic Fr = 18.533

Dunn's Multiple Comparisons Test

If the difference between rank sum means is greater than 13.117 then the P value is less than 0.05.

Comparison		Rank Sum Difference	P	value
Provenance vs.	NoProvenance	-16.000	*	P<0.05
Provenance vs.	NoQuality	-23.000	***	P<0.001
NoProvenance vs.	NoQuality	-7.000	ns	P>0.05

TABLE 7.34 TESTING RESULTS FOR HYPOTHESIS 4

Interpretation of testing results

The analysis indicates a difference in the ranking across the use of quality properties; the assessment considering provenance; and with no consideration of quality with $p < 0.0001$. For instance, the multiple comparisons indicate (at 0.05 and 0.001 level of significance) that the scores of the data sources with no consideration of provenance were greater than scores of data sources which were assessed by their ancestors. Note that the p-value for the Chi-square test is reported even though the sample size is small. In this case, the tabled value agrees with the Chi-square value.

7.8.3 Experiment 3

Initial conditions

Information System	Quality Criteria	Weights	Ranking Method	Condition	Granularity Level
TPCH	Accuracy	35	TOPSIS	1	Query
	Completeness	55		2	Database
	Timeliness	95			
	Uniqueness	25			
	Consistency	75			
	None	None		3	Database

TABLE 7.35 EXPERIMENT CONDITIONS FOR HYPOTHESIS 4

Testing results

Friedman Test (Nonparametric Repeated Measures ANOVA)

The P value is < 0.0001, considered extremely significant.
Variation among column medians is significantly greater than expected by chance.
The P value is approximate (from chi-square distribution) because exact calculations would have taken too long.
Calculation detail

Group	Sum of Ranks
Provenance	29.000
NoProvenance	53.000
NOQUALITY	80.000

Number of Rows = 27
Number of Columns = 3
Friedman Statistic Fr = 48.222
Dunn's Multiple Comparisons Test
If the difference between rank sum means is greater than 17.598 then the P value is less than 0.05.

Comparison		Rank Sum Difference	P value
Provenance vs.	NoProvenance	-24.000	** P<0.01
Provenance vs.	NOQUALITY	-51.000	*** P<0.001
NoProvenance vs.	NOQUALITY	-27.000	*** p<0.001

TABLE 7.36 TESTING RESULTS FOR HYPOTHESIS 4

Interpretation of testing results

The analysis indicates that there is a difference in the ranking across the use of quality properties; assessment considering provenance; and with no consideration of quality with $p < 0.0001$. For instance, the multiple comparisons indicate (at the 0.001 significance level) that the scores of the data sources with no consideration of provenance were greater than the scores of data sources which were assessed by their ancestors.

7.8.4 Conclusion

All the experiments have shown highly statistically significant changes in the ranking of queries and consequently in their outcomes when decision is made on the bases of either one of the following decision criteria: quality properties, assessment with data provenance, and no consideration of quality properties. We have enough evidence to accept the hypothesis 4.

"A Query outcome varies with respect to either one of the following decision criteria: quality properties, assessment with data provenance, and no consideration of quality properties."

7.9 Summary

The experimental hypotheses were set in terms of quality criteria, quality priorities, levels of granularity and data provenance to test if the qualitative information varies appropriately according with the specification of the context.

All the experiments conducted within this section have shown extremely significant evidence that our experimental hypotheses have been accepted with no possibility of being a result of chance. Therefore, we can conclude the following:

1. *The ranking of data sources changes according to the specification of quality criteria.*
2. *The ranking of data sources changes according to the specification of quality priorities.*
3. *The query outcomes change according to the assessment at different levels of granularity of their correspondent data sources.*
4. *Query outcomes vary with respect to a) the data sources it comes from b) no quality criteria consideration and c) quality criteria and against quality criteria along with data provenance.*

Having proved our experimental hypotheses, we can conclude that users are able to change the ranking of data sources and therefore query outcomes, using all the possible combinations of quality properties, quality priorities, and levels of granularity. This demonstrates that by changing the quality requirements data sources can be ordered in a different way, and that this variation is statistically significant.

7.10 Conclusions

We have demonstrated that the DQM prototype works in terms of functionality and it is able to help data consumers with any level of experience. The DQM provides qualitative information of derived data, and at database, relation, attribute levels of granularity within a multi-database environment (covered in section 7.2.2).

We have demonstrated that the DQM provides appropriated outcomes according with the quality of the data in section 7.2.3.

We have demostrate that the ranking of data sources correspond to the expected outcomes by changing the priority values of chosen quality criteria stated by the user (covered in section 7.2.4).

We have demonstrate that the prototype can provide qualitative information, which varies appropriately according to the context (covered in sections 7.3 to 7.9).

It would be reasonable to assume that by providing users with qualitative information, they could utilise such information to deal with extensional inconsistencies. However, more detailed and exhausted testing of the qualitative information been produced by the DQM needs to be carried out.

Chapter 8 Conclusions and Future Work

8.1 Review of the Thesis

This thesis was laid out in eight chapters. This section will state the main achievements of each chapter as well as the issues identified during the research process.

Chapter 1 Introduction:

The main motivation for the research was that users have no information by which to judge quality, e.g. whether data is correct, complete, or timely. When users query a database system, they get returned a set of data which is inherently presented as perfect, original, and atomic. However, data quality is an issue during data integration, users are facing difficulties to trust data when they face extensional inconsistencies since they have no qualitative information against which to consider those inconsistencies. Therefore, the research hypothesis was:

"It is possible to identify usable data quality criteria to measure and assess data quality of derived data and data at multiple levels of granularity. These can be enhanced by the use of provenance, and the qualitative measures can be used to derive a ranking of data sources based on the specification context by the users in a heterogeneous multi-database environment.

Consequently, we proposed a Data Quality Manager by establishing the following thesis objectives:

1. The identification from existing research of a set of usable and meaningful data properties as quality indicators to measure, assess and rank data sources, namely the Data Quality Reference Model (DQRM) (covered in section 3.4).

2. The identification of metrics to be used as data quality measurement instruments at database, relation, and attribute levels of granularity of primary data sources, namely the Data Quality Measurement Model (DQMM) (detailed in Section 3.5).

3. The identification of the processes required to represent, to interpret, and to assess data quality, namely the Data Quality Assessment Model, (DQAM) (see Chapter 3).

4. The implementation of a data provenance algorithm to help assessment processes for derived data sources (referred to in Chapter 4).

5. The identification and implementation of Multi-Attribute Decision Making methods to provide an overall quality score to rank data sources (covered in Chapter 5).

6. The design and development of a prototype as a proof of concept for direct user input of the query, quality properties and priorities (covered in Chapter 6).

7. Demonstrate that the prototype performs appropriately according to the specified requirements and can provide qualitative information, which varies appropriately according to the context. See Chapter 7 for further detail.

Chapter 2 Background:

Information Quality (IQ) in contrast to Data Quality (DQ) is concerned with data quality in context, and how the information is produced, and interpreted. Therefore, DQ should be considered before IQ to determine which data to trust, before determining the quality of the information generated for it.

From the literature review presented in the Background Chapter relative to our topic of interest, we identified the following issues:

- Current data management practice considers data as perfect, original, and atomic. The present research challenges the presumptions of perfection, primary authorship, and atomicity.

- There is no consideration of the process of integration (i.e. data fusion, data replication, or data transformation) during data quality measurement and assessment of derived data.

- Very few approaches have taken into account quality properties at attribute, data value and relation levels of granularity on databases, exceptions are [Scannapieco04b],[Naumann04].

- As we want to support a full range from experienced to naive users, only the assessment of process-criteria and objective-criteria rather than subjective-criteria should be considered because they can provide meanly and useful scores.

Chapter 3 The Data Quality Manager:

The chapter presented a conceptual framework for data integration where the DQM could help to deal with extensional inconsistencies.

This chapter discussed and identified the first three elements of the DQM, the Generic Data Quality Reference Model, which classifies and summarises a set of quality properties according to different user perspectives; the Measurement Model, which contains extended existing metrics for the assessment of primary data sources at database, relation and attribute levels of granularity; and the Assessment Model, which identifies a new granularity-oriented assessment classification (namely as direct and indirect assessment methods). However, there was an emerging issue:

- The direct and indirect assessment methods described in section 3.6 were not able to address derived data. This led us to the consideration of the assessment of derived data based on its origins.

The Data Quality Reference Model, the Measurement Model, and the Assessment Model for primary data sources correspond to the first three objectives of the research.

Chapter 4 Data Provenance

This chapter addressed the assessment of derived data by considering the quality indicators of its ancestors, through the extraction of data provenance. Therefore, the Data Quality Manager evolved to include data provenance as a mechanism to help data quality assessment and consequently resolve data inconsistencies between successor databases.

The assessment of derived data has been achieved in two ways:

- Assessment of derived data based on the quality properties of its ancestors, (covered in section 4.6).

- Assessment of derived data based on the aggregation of the quality properties of its ancestors (covered in section 4.7).

However, qualitative information is affected when derived data has been obtained from a fusion function. As the provenance algorithm does not trace queries, the quality score is not obtained automatically from the fusion function, but still provide a qualitative measure by the afore mentioned procedures.

We can conclude that the fusion function is not critical in terms that at this point the DQM can still assign a quality value to derived data and specially if it is able to use provenance to track down the fused elements

Having the assessment capability for derived data, we have achieved the objective 4 of this thesis, which correspond to the implementation of data provenance algorithm to help assessment processes for derived data sources.

We also detected the following issue:

Comparison between data sources by considering a number of associated quality properties. This issue is important in order to obtain the best outcome through the ranking of data sources.

Chapter 5 Multiple Attribute Decision Making (MADM) methods.

This chapter identified, selected, and explained a number of scaling and ranking methods from previous research.

The TOPSIS and SAW ranking methods developed in [Hwang81], and described in section 5.5, allowed the ranking of data sources through the comparison of multiple data quality dimensions, different weights, user quality priorities, at different levels of granularity.

- We proposed a number of pertinent combinations of scaling methods with the ranking methods according to the positive or negative direction of the quality criteria in order to obtain coherent, meaningful results.

- An experienced user should choose which ranking and scaling method is preferred. In the case of inexperienced users, the system will use the pertinent combination (from the recommendations made in Section 5.6) based on the characteristics of the quality criteria.

Emerging issues:

- The consideration of making available the selection of ranking and scaling methods or suggest to inexperienced users the best combination according to the characteristics of the quality properties.

- We require identifying which possible cases are available for the analysis of data quality according to the information available in the provenance metadata and quality metadata in order to implement the Data Quality Manager.

Chapter 6 Design and Implementation:

In Chapter 6, we described the design and the implementation of the prototype, according to the requirements explained through Chapters 3,4 and 5.

We believe that the Data Quality Manager can support a user context relative query across multiple databases. Where measures of data provenance along with a set of quality criteria would be used to determine quality of data sources at attribute, relation, and data source levels of granularity, and derived data.

The design and implementation of the DQM prototype with an appropriate level of functionality necessary to carry out the proof of concept corresponds to the fulfilment of objectives 5 and 6 described in Section 1.5.

Issues to address:

- At this point, we need to validate the DQM prototype against the specification of the model, and to verify that the Data Quality Manager (DQM) can provide appropriate information about the qualitative nature of the data been returned from the data sources.

Chapter 7 Testing and Experimentation:

The objectives of the testing plan were to determine if the prototype works in terms of achievement of requirements, and the appropriateness of the quality information synthesized.

From the testing plan, we conclude that:

- The prototype is addressing expert users by providing a high level of detail in the specification of the context and the functionality of the system.

- The prototype suggests a combination of ranking and scaling methods depending on the positive or negative direction of the quality properties.

- The prototype allows users the analysis of data quality and ranks data sources based on a desired context at different levels of granularity under three identified based test cases.

- The quality estimation changes on the basis of the granularity level under the same query event.

- We have demonstrated that the DQM provides appropriated outcomes according with the quality of the data in section 7.2.3.

- We have demonstrated that the DQM provides appropriated ranking of data sources outcomes according with the quality of the data in section 7.2.4.

In the second part of the chapter, we designed and conducted an experimentation plan on inferential and descriptive statistics to demonstrate that the DQM can provide qualitative information, which varies appropriately according to the context specified by the user.

The implementation and testing of the DQM resulted in a richer body of information. Therefore, a number of experimental hypotheses were set in terms of the combinations of quality criteria, priorities, levels of granularity and data provenance. The following experimental hypotheses were proved as a set of sanity checks.

1. The ranking of data sources changes according to the specification of quality criteria.

2. The ranking of data sources changes according to the specification of quality priorities.

3. The query outcomes change according to the assessment at different levels of granularity of their corresponding data sources.

4. Query outcomes vary with respect to a) the quality of the data sources it comes from b) the no consideration of data quality and c) the consideration of data quality assessed by data provenance.

The set of proven experimental hypotheses demonstrate that by changing the quality requirements, the data sources can be ordered in a different way, and that this variation is statistically significant.

The outcome of Chapter 7 was to demonstrate that the prototype performs appropriately according to the specified requirements, and can provide qualitative information, which varies appropriately according to the context. Such outcome corresponds to the achievement of the last objective of this work, presented as objective 7 in Section 1.5.

8.2 Contributions to Research

8.2.1 Framework for data integration considering data quality

We developed a framework for data integration considering data quality. The framework has been built on a large body of work that had been done by others [Wang96], [Naumann00], [Gertz04]. However, this framework has considered the use of data quality with known provenance within the data integration process, which is a novel approach.

8.2.2 The Data Quality Manager

The DQM architecture is novel, each element has been implicitly used before, but in an isolated mode. It proposes generic and expressive models for a set of quality properties, measurement, and assessment of data quality. Each model provides novel elements, which are mentioned in the following sections.

The DQM provides qualitative information to any level of experience users to extend the scope and range of information available relative to the query they have presented within the quality properties and priorities they state.

8.2.3 The Data Quality Reference Model

The DQRM contains a set of data quality properties classified and summarized according to different user perspectives such as internal and external focuses or representation, value, and context. Previous classifications of data quality do not consider uniqueness as a data quality criterion, the DQRM includes this quality property, because duplicate values are a very well know issue within Database

Management Systems, and in case of a multi-database environment this problem might be a cause of extensional inconsistencies.

8.2.4 The Data Quality Measurement Model

This model assembles already existing data quality metrics e.g. [Ballou98], [Motro98], [Pipino02], [Naumann03], and extends these metrics for the measurement at database, relation, tuple and attribute levels of granularity following the example for completeness given in [Scannapieco04b]. Within the Measurement Model, there is an identification of data quality properties such as accuracy, uniqueness, currency, timeliness, volatility, and response time. From our research we can state that there are no other systems that use combinations of several quality properties at different levels of granularity as we are using with the Data Quality Measurement Model.

8.2.5 The Data Quality Assessment Model

The DQAM identifies and includes temporality, positive and negative directions from existing research e.g. [Burgess03b],[Naumann02]. As this project was implemented at data value level, the quality properties were assessed through query processing such as parsing, sampling or continuous assessment [Naumann00]. Hence, the scores were represented in a quantitative manner.

The Assessment Model contributes with the following novel elements:

- **Data Quality Assessment by considering data provenance.**

The Data Quality Assessment Model provides a mechanism for tracking data provenance for the assessment of quality of derived data. Previous approaches work from the presumption of primary authorship and the presumption of atomicity. Therefore, the introduction of data provenance as a mechanism of considering qualitative information of derived data is novel.

- **The Assessment of derived data based on the quality properties of its ancestors**

With the description of provenance, users are able to trace back the quality properties of any data by selecting each data ancestor. Therefore to have enough information to trust or not to trust the data. In summary, users are able to compare data sources, and to trust one data source against another by comparing the quality properties of their ancestors.

- **The assessment of derived data based on the aggregation of the quality properties of its ancestors**

In this case, the DQM is able to assign quality scores to derived data by the aggregation of the quality properties of its ancestors. This assessment requires that all the quality scores of the corresponding ancestors are available.

- **The granularity-based assessment clasification**

The following assessment classification is according to the level of granularity in which the assessment is carried out, and it is novel.

a) Direct assessment: The process of assessment relates directly to the level of granularity such as uniqueness, which relates at the tuple level.

b) Indirect assessment: The score is calculated based on other scores at other levels of granularity of the same source, such as accuracy at the relation level which value depends on accuracy at the row level.

c) Assessment by provenance: The score of an object is computed based on the quality indicators of its ancestors.

8.2.6 The Ranking of Data Sources

The methods TOPSIS and SAW have been already utilised for the ranking of data sources by F. Naumann in [Naumann02] and M. Burgess in [Burgess02].

According to our analysis regarding the possible combinations of scaling methods with the ranking methods, the recommendation to obtain coherent and meaningful results (and therefore a more reliable decision) with both ranking methods, has been to use TOPSIS with Vector Normalization or SAW with Linear Scale Transformation in case positive and negative criteria are involved. In case all criteria are positives or negatives, the Vector Normalization method is the best option with SAW. As far as we know, the previous analysis and recommendation have not been done before.

8.2.7 Data Quality Model within a multi-database environment.

We have now a developed a system that provides representation, measurement and assessment of data quality not previously available within a multi-database environment

8.2.8 Provision of a facility to profile users in terms of the context of their query

By default, the DQM will provide:

- An appropriate combination of scaling with ranking methods.
- A stereotypical criteria according to the type of user. At the two extremes: in the case of naive users, the DQM provides the context automatically; in the case of expert users, the user will have the ability to define everything: scaling, ranking, quality properties and the priorities for a higher level of analysis.

8.3 Conclusions

This thesis has considered the association of data quality measures with data retrieved during the querying process on the basis of a set of data sources, a set of data quality properties, and their corresponding priorities.

Databases have been considered as perfect, this research assumes that data are of different and variable quality, challenging the presumption of perfection and providing measures of data quality.

Nowadays, organizations are facing a huge amount of data sources available for query. Consequently, we cannot assume that data sources are the original point of the data they are using. Therefore, discarding the presumption of primary authorship, we have used data provenance as a mechanism to deal with the origination of data within data sources to either assure the user that the data has been provided from the primary source or show evidence of the data provenance and associated quality.

Rather than assuming that data is atomic we can now either assure the user the data is atomic or it is composed data and demonstrate what the atomic values from which it was generated are and their associated quality.

We have identified a set of criteria for assessment of data quality, suitable to support a range of users from naive users to highly experienced users and then to provide user-independent quality scores.

We propose to use the quality of measures as the means of resolving inconsistencies between the same data in different databases.

Data has been fused, replicated, and transformed to resolve intensional inconsistencies, degrading data quality consequently. Therefore, data provenance along with the quality criteria allows the assessment at different levels of granularity, estimates the quality of derived data, and ranks them at the data value level.

We have built a Data Quality tool that allows users to rank query data sources, and execution plans at multiple levels of granularity.

We have validated the functionality and capability of the DQM prototype against the specifications.

Based on the testing plan we have demonstrated that the DQM prototype does what it is required to do in terms of functionality.

The prototype provides appropriate scores according with the expected outcomes based on the actual quality of data.

The ranking of the data sources correspond to the expected outcomes by changing the priority values of chosen quality criteria stated by the user.

We have shown that the DQM can estimate an overall score between data sources by providing data quality information at different levels of granularity, which can vary according to the context specification of data consumers in section.

We can conclude that it has been possible to identify usable data quality criteria to measure, and assess data quality of primary data sources at multiple levels of granularity, and derived data. Such quality information was enhanced by the use of provenance, and the qualitative measures could be used to derive ranking of data sources based on the specification context by the users utilising this known criteria all within an heterogeneous multi-database environment.

8.4 Limitations

8.4.1 Metadata maintenance

For each data source involved in the federation, we required to capture metadata, for the data covering quality scores and data provenance information. The implicit management of metadata is a typical restriction of all metadata-based systems.

8.4.2 Measurement Model

The measurement model is restricted to relational databases because most of the metrics we obtained from previous research, and they were directly related to the relational data model. However, it can be extended for the inclusion of XML documents and semi-structured data. For example, the Sybase Adaptive Server provides the facility either to interact with XML by managing XML documents and to store them in a relational database (which is not recommended), or to generate XML from relational databases.

As we mentioned in Chapter 3, the Measurement Model contains quality properties at data value level only for a user independent measurement. However, it could be extended by considering all the quality properties mentioned in the Generic Data Quality Reference Model. Such measures could then be stored as part of the user profile.

8.4.3 Assessment Model

Regarding the assessment of derived data based on the aggregation of the quality scores of its ancestors, we have utilised illustrative aggregation functions only. The rationale for using a pesimistic or optimistic approach for the aggregation functions is that all the data should be treated in the same way according to the application requirements. The key point is the consistency in using the aggregation functions.

In the case of fused data, the provenance algorithm is not able to obtained the elements involved in the fusion function to aggregate its quality scores and obtain the quality score of fused data automatically. Therefore, the alternative is to assess it just in terms of the quality properties of its fused elements, which can be obtained from the where provenance algorithm through the query that produces the derived data.

8.5 Future work

8.5.1 Fusion function

In the case of fused data, the fusion function is important for improving the accuracy of the qualitative information already estimated. Therefore, the provenance algorithm should be enhanced in order to trace queries and obtain the elements of the data fusion

along with the fusion function to aggregate the corresponding quality measures according to the fusion function.

The fusion function used to derived data should be considered when comparing fused data. Therefore, it shall be included as an independent variable within the experiments carried out in Chapter 7. Therefore, the next stage will be to exhaustively test as part of future work.

8.5.2 Aggregation of quality scores

The assessment of derived data based on the aggregation of the quality scores of its corresponding ancestors was illustrative. The analysis of which aggregation functions are appropriate corresponds to future research.

8.5.3 Prototype

We have built a system prototype considering data as, primary sources; derived data with no quality scores for it, but with enough quality scores of its ancestors; and derived data with enough information to compute its quality scores (see Section 6.2.6). There is a forth scenario, which is a combination of the previous scenarios. The case of derived data with incomplete data provenance, and incomplete quality scores it. For a prototype, it is not necessary to address the forth scenario, because as long as the three basic scenarios are available, the solution is trivial. This extension is subject to future work.

Having obtained sensible expected outcomes, we can conclude that the prototype behaves as is expected in terms of the accuracy of the answers it produces. However, these tests are at indicative level only.

8.5.4 User Stereotypes

We have proposed user stereotypes as a mechanism that we would utilize to present to novice users who require higher levels of automation of the specification of the context of the query. Stereotyping is something at an early stage of development. We have identified some simple stereotypes very briefly that can be referred to as useful. See Appendix D for further detail. Therefore, the identification of user stereotypes is part of a future work.

8.5.5 Information Quality

The outcome of this work could fit in the future work on Information Quality (IQ) because we have discarded the presumptions of perfection, primary authorship, and atomicity. Therefore, IQ measures could be developed from the Data Quality (DQ) assessment methods we have identified.

References

[Angeles04] P. Angeles and L.M. MacKinnon, *"Detection and Resolution of Data Inconsistencies, and Data Integration using Data Quality Criteria"*, Proceedings of QUATIC 2004: Conference for Quality in Information and Communications Technology .Instituto Portugues da Qualidade, ed., pp. 87-94., ISBN 972-763-069-3, Porto, Portugal, 2004.

[Angeles05] P. Angeles and L.M. MacKinnon, *"Tracking Data Provenance with a Shared Metadata"*, Proceedings of PREP 2005: Postgraduate Research Conference in Electronics, Photonics, Communications and Networks, and Computing Science, pp. 120-121, Lancaster England, U.K., 2005.

[Angeles05b] P. Angeles and L.M. MacKinnon, *"Quality Measurement and Assessment Models Including Data Provenance to Grade Data Sources"*, International Conference on Computer Science and Information Systems" at the Athens Institute for Education and Research, pp. 101-118, Athens, Greece, 2005,.

[Anokhin01] P. Anokhin and A. Motro, *"Data Integration: Inconsistency Detection and Resolution Based on Source Properties"*, Proc. of FMII 2001, 10th International Workshop on Foundations of Models for Information Integration. Viterbo, Italy., 2001

[Anokhin03] P. Anokhin and A. Motro, *"Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources"*, Technical Report ISE-TR-03-06, Information and Software Engineering Dept., George Mason Univ., Fairfax, Virginia, 2003.

[Baldoni03] R. Baldoni, M. Contenti, A. Virgillito. The Evolution of Publish/Subscribe Communication Systems "Future Directions of Distributed Computing", Springer Verlag LNCS Vol. 2584, 2003

[Ballou98] D. Ballou, G. Tayi and Guest Editors, "Examining Data Quality", *Communications of the ACM*, vol. 41,no.2, pp.54-57, 1998.

[Ballou98b] Ballou, D.P., Wang, R.Y., Pazer, H. and Tayi, G.K. "Modeling information manufacturing systems to determine information product quality". Management Science 44, 4 April,1998, 462–484.

[Barnett84] Barnett, V. and Lewis, T.: 1984, *Outliers in Statistical Data*, John Wiley & Sons, New York.

[Batini86] C. Batini, M. Lenzerini and S.B. Navathe *"A comparative Analysis of Methodologies for Database Schema Integration"*, ACM Computing Surveys, vol. 18, no. 4, pp. 323-364, 1986.

[Bertolazzi03] P.Bertolazzi, L.De Santis, M.Scannapieco, "Automatic Record Matching in Cooperative Information Systems", Proceedings of the ICDT 2003 International Workshop "Data Quality in Cooperative Information Systems" (DQCIS 2003), Siena, Italy, 2003

[Bhagwat04] D. Bhagwat, L. Chiticariu, W. Tan, and G. Vijayvargiya. An Annotation Management System for Relational Databases. In Proceedings of the International Conference on Very Large Data Bases (VLDB), pp 900-911,Toronto, Canada, 2004.

[Bock98] R.K. Bock and W. Krischer,"The Data Analysis BriefBook" Springer 1998.

[Bovee01] Bovee M.,Mark B., Srivastava E.P., "A Conceptual Framework and Belief Function Approach to Assessing Overall Information Quality". Proceedings of the Sixth International Conference on Information Quality 2001.

[Buchheit02] Buchheit R. B., "Vacuum: Automated Procedures for Assessing and Cleansing Civil Infrastructure Data", PhD Thesis, May 2002

[Buneman98] Buneman P., Liberman M., Overton C.J., Tannen V.,"Data Provenance", pp. 1-16, homepage: http://www.cis.upenn.edu/~wctan/DataProvenance,1998

[Buneman00] P. Buneman S. Khanna W. Tan, "Data Provenance: Some Basic Issues", Foundations of Software Technology and Theoretical Computer Science 2000.

[Buneman01] P. Buneman, S. Khanna, and W. Tan, "Why and Where: A Characterization of Data Provenance". Proceedings of the International conference on Database Theory (ICDT), Springer ed., pp. 316-330,London, United Kingdom, 2001.

[Buneman02] Buneman P., Khanna S., Tan W.,Tajima K. (2002), "Archiving Scientific Data", Proceedings of ACM SIGMOD International Conference Management of Data., pp.1-12

[Buneman04] P. Buneman, M. Liberman, C.J. Overton and V. Tannen, *"Data Provenance"*, http://www.cis.upenn.edu/~wctan/DataProvenance, [(date information as accessed by the author citing the references, e.g. 17 Aug. 2004.)]

[Burgess02] M.S.E. Burgess, W. A. Gray, and N.J. Fiddian, "Establishing a Taxonomy of Quality for Use in Information Filtering", Proceedings of the 19th British National Conference on Databases (BNCOD 2002), Lecture Notes in Computer Science: Advances in Databases (LNCS 2405), Sheffield, UK, July 2002, pp 103-113.

[Burgess03] M.S.E. Burgess, W.A. Gray, and N.J. Fiddian, "A Flexible Quality Framework For Use Within Information Retrieval", Proceedings of the 8th International Conference on Information Quality (ICIQ-03), Cambridge, MA, USA, November 2003.

[Burgess03b] M.S.E. Burgess, "Using Multiple Quality Criteria to Focus Information Search Results", PhD Thesis, September 2003.

[Burgess04] M.S.E. Burgess, W.A. Gray, and N.J. Fiddian. "Quality Measures and the Information Consumer", Proceedings of the 9th International Conference on Information Quality (ICIQ-04), Cambridge, MA, USA, November 2004.

[Cappiello02] C. Cappiello, C. Francalanci, P. Missier, B. Pernici, P. Plebani, M. Scannapieco, A. Virgillito, "Presentation of Metadata and of the Quality Certificate", Dicembre 2002, DaQuinCIS Project Report (http://www.dis.uniroma1.it/~dq/docs.html)

[Cappiello03] C.Cappiello, C.Francalanci, B.Pernici, P.Plebani, M.Scannapieco, "Data Quality Assurance in Cooperative Information Systems: a Multi-dimension Quality Certificate". Proceedings of the ICDT 2003 International Workshop "Data Quality in Cooperative Information Systems" (DQCIS 2003), Siena, Italy, 2003

[Charnes78] A. Charnes, W. Cooper, and E. Rhodes. "Measuring the efficiency of decision making units", *European Journal of Operational Research*, pp. 429-444, 1978.

[Cavano78] Cavano J."A Framewok for the Measurement of Sotware Quality",Rome Air Development Center, James A. McCall, General Electric Company (1978),pp.133-

139.

[Codd70] C. F. Codd, "(1970), "A Relational Model for Large Shared Data Banks" *Communications of the ACM*, Vol. 13, No. 6, pp. 377-387.

[Codd79] C. F. Codd, "(1979), "Extending the Database Relational to Capture More Meaning" ACM Trans. On Database Systems, 4 pp. 262-296.

[Codd86] C. F. Codd, "(1986), "Missing Information (Applicable and Inapplicable) in Relational Databases", ACM SIGMOD Record, Vol. 15(4).

[Codd90] C. F. Codd, (1990) "The Relational Model for DBM version 2" Adisson Wesley.

[Cui00] Y.Cui, J.Widom, "Practical lineage tracing in data warehouses", in International Conference on Data Engineering (ICDE), pp 367-378, 2000.

[Dataflux] DataFlux a SAS Company home page located at: http://www.dataflux.com/, [(date information as accessed by the author citing the references, 05-September-2006.)]

[DataFluxv7] DataFlux Corporation, DataFlux Version 7 Tecnology, The Convergence of Data Quality and Data Integration located at: http://www.dataflux.com/, [(date information as accessed by the author citing the references, 05-September-2006.)]

[Date89] Date, C.J. ,(1989) "The Default values approach to missing information" Relational Database Writings 1989-1991,pp 343-354

[Date89b] Date, C.J. ,(1989) "Be Careful with SQL EXISTS" Database Programming and Design vol.2(9), pp 50-52

[Date04] Date, C.J. ,(2004) "An Introduction to Database Systems" 8th. Edition Pearson Education Inc.

[El-Khatib00] H.T. El-Khatib, H. Williams, L.M. MacKinnon and D.H. Marwick, "A framework and test-suite for assessing approaches to resolving heterogeneity in distributed databases", Information and Software Technology Vol. 42, No. 7, pp. 505-515, 2000.

[Evoke] Evoke Software Corporation and Informatica Data Quality, "Empowering Business Professionals to Manage Data Quality Across the Enterprise", located at http://www.informatica.com/products/data_quality/default.htm/, [(date information as accessed by the author citing the references, 05-September-2006.)]

[Fensel05] Fensel D.,"Information Integration with Ontologies", John Wiley and Sons, Technology & Industrial Arts, 196 pages, ISBN 0470010487, 2005.

[Gertz98] M. Gertz and I. Schmitt, "Data Integration Techniques Based on Data Quality Aspects", *3rd National Workshop on Federal Databases*, Magdeburg, Germany, 1998.

[Gertz98b] M. Gertz, "Managing Data Quality and Integrity in Federated Databases", *Second Annual IFIP TC-11 WG 11.5 Working Conference on Integrity and Internal Control in Information Systems*. Warrenton, Virginia, Kluwer Academic Publishers, 1998.

[Gertz04] M. Gertz, "Report on the Daugstuhl Seminar, Data Quality on the Web", *SIGMOD Record*, Vol. 33, No. 1, Mar. 2004.

[GraphPad] GraphPad Software located at http://www.graphpad.com/instat/instat.htm [(date information as accessed by the author citing the references, 05-September-2006.)]

[GraphPad03] H. Motulsky, "The InStat Guide to choosing and interpreting statistical tests", GraphPad InStat Software version 3. http://www.graphpad.com/Downloads/InStat3.pdf [(date information as accessed by the author citing the references, 05-September-2006.)]

[Gravetter00] Gravetter F.J., Wallnau L.B.,"Statistics for the Behavioral Sciences", 5th Edition, State University of New York, Brockport , ISBN 0534359264, 2000.

[Hernandez98] Hernandez, M. A., and Stolfo, S. J. Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery 2, 1 (1998), 9–37.

[Hwang81] C.L. Hwang and K. Yoon, "Multiple Attribute Decision Making: Methods and Applications: a state-of-the-art survey", Berlin; Springer-Verlag.

[Hwang95] Hwang C.-L. and Yoon K. (1995, "Multiple Attribute Decision Making An Introduction" Series: Quantitative Applications in the Social Sciences, Volume # 104, Sage Publications Inc.

[Jarke98] M. Jarke, M.A. Jeusfeld, C. Quix, P.Vassiliadis "Architecture and Quality in Dta Warehouses an extended Repository Approach", Journal on Information Systems, Vol 24",No.3,pp.229-253"1999", URL "citeseer.ist.psu.edu/jarke99architecture.html"

[Kahn02] Beverly K. Kahn, Diane M. Strong, and Richard Y. Wang, "Information Quality Benchmarks: Product and Service Performance" April 2002/Vol. 45, No. 4ve Communications of the ACM pp.184-192

[Kano84] Noriaki Kano, Seraku, Takahashi, and Tsuji, Hinshitsu,"Attractive Quality and Must-be Quality", Vol. 14, No. 2, 1984.

[Karr05] Karr A.F. Sanil A.OP. Banks D.L.," Data Quality: A Statistical Perspective", Technical Report 151, March 2005, National Institute of Statistical Sciences.

[Kon95] H. Kon , E. Madrick, and M. Siegel, "Good answers from bad data", Sloan WP #3868, 1995.

[Lee04] Lee Y. and Strong D. "Knowing-Why about Data Processes and Data Quality", Journal of Management Information Systems, Vol. 20, No. 3, pp. 13 – 39. 2004.

[Lesser00] U. Leser and F. Naumann, "Query Planning with Information Quality Bounds", *Proceedings of the 4th International Conference on Flexible Query Answering, (FQAS00)*, Warsaw Poland, 2000.

[Linthicum99] Linthicum D.S., "Enterprise Application Integration", Addison-Wesley Professional, Computers / Languages / Programming, 377 pp, ISBN 0201615835, 1990.

[Lipschutz02] P. Lipschutz, M. Lipson, "Easy Outline Linear Algebra" Schaum (Schaum's Easy Outlines), 144 pp. 2000.

[Loshin05] Loshin M., "Developing Information Quality Metrics", Data Modeling Review Magazine May 2005 http://www.dmreview.com/article_sub.cfm?articleId=1026061

[Loshin05b] Loshin, D. "Data Provenance and Data Quality", DataFlux Community of Experts, located at: http://www.dataflux.com/blog/archives/2005/12/13/data-

provenance-and-data-quality/ , [(date information as accessed by the author citing the references, 05-September-2006.)]

[Little87] Little, R. J. Statistical Analysis with Missing Data. Wiley, New York, 1987.

[Maletic00] Maletic, J. I., and Marcus, A. Data cleansing: Beyond integrity checking. In Proceedings of the Conference on Information Quality (IQ2000) (Massachusetts Institute of Technology, October 2000), pp. 200–209.

[Marchetti03] C.Marchetti, M.Mecella, M.Scannapieco, A.Virgillito, R.Baldoni, "Data Quality Notification in Cooperative Information Systems", Proceedings of the ICDT 2003 International Workshop "Data Quality in Cooperative Information Systems" (DQCIS 2003), Siena, Italy, 2003.

[Marchetti03-2] C. Marchetti, M. Mecella, M. Scannapieco, A. Virgillito. Enabling Data Quality Notification in Cooperative Information Systems through a Web-Service Based Architecture (short paper) Proceedings of the 4th International Conference on Web Information Systems Engineering, Roma, Italy, December 2003

[Mecella02] M. Mecella , M. Scannapieco, A. Virgillito, R. Baldoni , T. Catarci, C. Batini, "Managing Data Quality in Cooperative Information Systems" Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002

[Mecella02-2] M. Mecella, M. Scannapieco, A.Virgillito, R. Baldoni, T. Catarci and C. Batini, "Managing Data Quality in Cooperative Information Systems". In Proceedings of the Tenth International Conference on Cooperative Information Systems (CoopIS 2002), Irvine, CA, 2002.

[Mecella03] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, C. Batini. Managing Data Quality in Cooperative Information Systems, Journal of Data Semantics, Volume I, LNCS 2800, 2003

[MacKinnon98] L.M. MacKinnon, D.H. Marwick D.H., and H. Williams., "A Model for Query Decomposition and Answer Construction in Heterogeneous Distributed Database Systems", Journal of Intelligent Information Systems, pp. 69-87. 1998.

[Missier03] P. Missier, C.Batini, A Multidimensional model for Information Quality in Cooperative Information Systems , Procs. 8th International Conference on Information Quality, ICIQ 2003, Cambridge, Ma.

[Motro98] A. Motro and I. Rakov I, "Estimating the Quality of Databases", *Proceedings of FQAS 98: Third International Conference on Flexible Query Answering Systems*, T. Andreasen, H. Christiansen, and H.L. Larsen, ed., pp. 298-307. Roskilde, Den.mark, Springer-Verlag, Berlin, Germany, 1998.

[Motulsky95] H. Motulsky. "Intuitive Biostatistics", *Oxford University Press*, September, 1995, ISBN 0195086074.

[Naumann98] F. Naumann, "Data Fusion and Data Quality", *Proceedings of the New Techniques & Technologies for Statistics Seminar.* Surrent, Italy 1998.

[Naumann99] F. Naumann, "Quality-driven Integration of Heterogeneous Information Systems", *Proceedings of the 25th Very Large Data Bases Conference (VLDB99)*, Edinburgh, Scotland, 1999.

[Naumann99-2] F. Naumann and C. Roker, "Do Metadata Models meet IQ Requirements", *Proceedings of the International Conference on Information Quality*, MIT Cambridge, 1999.

[Naumann00] F. Naumann and C. Roker C., "Assessment Methods for Information Quality Criteria", *Proceedings of the International Conference on Information Quality (IQ2000)*, Cambridge, Mass., 2000.

[Naumann01] F. Naumann, "From Databases to Information Systems-Information Quality Makes the Difference", *Proceedings of the International Conference on Information Quality (IQ2001)*, Cambridge, Mass., 2001.

[Naumann02] F. Naumann, "Quality-Driven Query Answering for Integrated Information Systems", *Lecture Notes in Computer Sciences LNCS 2261*, Springer Verlag, Heidelberg, 2002.

[Naumann02-2] F. Naumann and M. Haeussler, "Declarative Data Merging with Conflict Resolution", *Proceedings of the International Conference on Information Quality (IQ2002)* Cambridge, Mass., 2002.

[Naumann03] F. Naumann, J. Freytag and U. Lesser, "Completeness of Information Sources", *Workshop on Data Quality in Cooperative Information Systems (DQCIS2003)*, Cambridge, Mass., 2003.

[NISS] The National Institute of Statistical S. website, http://www.niss.org/index.html, [(date information as accessed by the author citing the references, e.g. 15 Apr. 2006.)]

[Pipino02] L. Pipino, W.L. Yang and R. Wang, "Data Quality Assessment", *Communications of the ACM*,vol. 44 no. 4e, pp.211-218, 2002.

[Parsian99] A. Parssian, S. Sumit and V. Jacob, "Assessing Data Quality for Information Products", *Proceeding of the 20th International Conference in Information Systems (ICIS1999)*, Charlotte, North Carolina USA, pp. 428-433, 1999.

[Pierce04] E. Pierce, "Assessing Data Quality with Control Matrices", *Communications of the ACM*, vol.47, no. 2, pp.82-86, 2004.

[QOD-12-24-03] Boone County Clerk L. Garofolo, "I about had a heart attack" Quote of the Day, Information Week Online, December 24, 2003.

[Rajola03] Rajola F., "Customer Relationship Management Organizational and Technological Perspectives", Springer, 172 pages, ISBN: 978-3-540-44001-7, 2003.

[Redman96] Redman "Data Quality for the Information Age", Boston, MA., London : Artech House, 1996.

[Reiter78] Reiter, R., "On Closed World Databases", Logic and Databases (H.Gallaire and J. Minker, eds.), PlenumPress, 1978.

[Scannapieco02] M. Scannapieco, T. Catarci, *"Data Quality under the Computer Science Perspective"*. Journal of "Archivi & Computer", Volume 2, 2002.

[Scannapieco04] M. Scannapieco, A. Virgillito, M. Marchetti, M. Mecella, R. Baldoni. "The DaQuinCIS Architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems" . To appear on Information Systems, Elsevier, pp. 551-582, 2004.

[Scannapieco04b] M. Scannapieco, C. Batini, "Completeness in the Relational Model: A Comprehensive Framework", Research Paper, in Proceedings of the 9h International Conference on Information Quality (ICIQ-04, Cambridge, MA, USA, November 2004.

[Scannapieco05] M. Scannapieco, P. Missier, C. Batini: Data Quality at a Glance. Datenbank-Spektrum, vol. 14, 2005.

[Sheskin03] Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition, *Western Connecticut State University, Danbury, USA*

[Sheth90] A. Sheth and L. Larson, "Federated Database Systems for Managing Distributed Heterogeneous and Autonomous Databases", *ACM Computing Surveys*, vol. 22, no. 3, pp.184-236, 1990.

[Sheth92] Sheth A., Kashyap V. (1992), "So Far (Schematically) yet So Far (Semantically)", Proceedings of the IFIP DS-5 Conference on Semantics of Interoperable Database Systems, Elsevier Publisher., pp. 283-312.

[Shtub99] Shtub A., "Enterprise Resource Planning (ERP) The Dynamics of Operations Management", Springer, 168 pages, ISBN 0792384385, 1999.

[Strong97] D.M. Strong, W.L. Yang and R.Y. Wang, "Data Quality in Context", *Communications of the ACM*, vol. 40, no. 5, pp.103-110, 1997.

[Strong97-2] D.M. Strong, W.L. Yang and R.Y. Wang, "10 Potholes in the Road to Information Quality", *Proceedings of IEEE*, vol.18, no. 9162, pp.38-46, 1997.

[Tan04] W.C. Tang, "Research Problems in Data Provenance" Bulletin of the IEEE Computer Society Technical Committee on Data Engineering Vol. 27 No. 4, pp. 45-52. 2004.

[TDQM] The MIT Total Data Quality Management web site, http://web.mit.edu/tdqm/, [(date information as accessed by the author citing the references, e.g. 15 Apr. 2006.)]

[TPC] Transaction Processing Performance Council (TPC) www.tpc.org, info@tpc.org

[TPC-C] TPC Benchmark ™ C Standard Specification Revision 5.7 April 2006, Transaction Processing Performance Council (TPC) www.tpc.org, info@tpc.org © 2006 located at: http://www.tpc.org/tpcc/spec/tpcc_current.doc

[TPC-H] TPC Benchmark ™ H (Decision Support), Standard Specification Revision 2.3.0 Transaction Processing Performance Council (TPC) www.tpc.org, info@tpc.org © 2006, located at http://www.tpc.org/tpch/spec/tpch2.3.0.pdf

[TPC-H-2] C. Ballinger, "Relevance of the TPC-D Benchmark Queries: The Question You Ask Every Day ",*NCR Parallel Systems*,located at ttp://www.tpc.org/information/other/articles/TPCDart_0197.asp

[Transact02] Sybase, Inc, "Transact-SQL® User's Guide", Adaptive Server Enterprise12.5, Document ID: 32300-01-1250-03, Last Revised: August 2002. website locatedat:http://manuals.sybase.com/onlinebooks/groupas/asg1250e/sqlug/@Generic__ BookView, [(date information as accessed by the author citing the references, 15-March-2007.)]

[Trillium] Trillium® Protocol Software, website located at: http://www.dataflux.com/blog/archives/2005/12/13/data-provenance-and-data-quality/ , [(date information as accessed by the author citing the references, 05-September-2006.)]

[Tsichritzis82] Tsichritzis D.C. and Lochovsky F.H., "Data Models", Prentice-Hall, ISBN:0131964283, 1982.

[Ullman84] Ullman, J.D , "On the Foundations of the Universal Relation Model", ACM Trans. on Database Systems, 9(2), pp.283–308, 1984

[Wand96] Y. Wand and R. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations", *Communications of the ACM*, vol. 39, no. 11, pp.86-95, 1996.

[Wang93] R.Y. Wang, M.P. Reedy, and A. Gupta, "An Object-Oriented Implementation of Quality Data Products". *Workshop on Information Technology Systems, O. Florida, 1993.*

[Wang95] R.Y. Wang, V.C. Storey, and C.P. Firth, "A Framework for Analysis of Data Quality Research," *IEEE Trans. Knowledge and Data Eng.*1995, pp. 623-640

[Wang96] Wang R. Y., Strong D.M. *"Beyond accuracy: What data quality means to Data Consumers"*, Journal of Management of Information Systems, vol. 12, no 4 1996, pp. 5 -33].

[Wang98] R. Wang, "A Product Perspective on Total Data Quality Management", *Communications of the ACM*, vol. 41, no. 2, pp.58-65, 1998.

[Wang02] Wang R.Y, Ziad M. and Lee Y., "Data Quality", The Kluwer International Series on Advances in Database Systems. Kluwer Academic Publishers. 2001.

[Wilcoxon64] F. Wilcoxon, R.A. Wilcoxon, "Some Rapid Approximate Statistical Procedure", 1964.

[Woodruff97] A. Woodruff, M. Stonebraker. "Supporting fine-grained data lineage in a database visualization environment", in the International Conference on Data Engineering (ICDE), pp 91-102, 1997.

[Yang02] L. Yang, D. Strong and R. Wang, "AIMQ: A Methodology for Information Quality Assessment", *Information and Management,* vol. 40, no. 2, pp. 133-146, 2002.

[Zhang04] W. Zhang. "Handover Decision Using Fuzzy MADM in Heterogeneous Networks". *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2004)*, Atlanta, 2004.

Appendix A Data Quality Concepts

This appendix contains the concepts of the Reference Model identified in Chapter 3.

Ability to represent nulls	The ability of the format to provide a good form of nulls representation.
Accessibility	The extent to which data is accessible in terms of availability and security and cost, data might be available but inaccessible for security purposes; data might be available but expensive.
Accuracy	The measure of the degree of agreement between a data value o collection of data values and source agreed to be correct in [Lee04], the metric is defined as the ratio between the correct values and the total values in the data source.
Amount of data	The extent to which the volume of data is appropriate for the task at hand.
Appropriateness	A format is more appropriate than other in terms if it is better suited to user's needs [Redman96]
Assessment	A systematic process of collecting and analysing data to determine the status of data quality.
Availability	The extent to which data is available, or easily and quickly retrievable.
Believability	Is the extent to which data is accepted or trusted as correct by the user without verification. When it is accepted as true, real and credible.
Completeness	The extent to which data is not missing [Redman96], [Pipino02], it is divided by two quality dimensions coverage, and density in [Naumann03].
Cost	The cost associated with providing poor quality data.
Coverage	Measure for the number of tuples a source stores. Probability that an entity of the world is represented in the source. [Naumann03]

Currency	Time interval between the latest update of a data value and the time it is used [Wang93], [Motro98].
Density	Measure for how well the attributes stored at a source are filled with actual (non-null) values [Naumann03].
Extensional inconsistencies	Data value differences between the participating data sources during data integration [Angeles04], [Motro98].
Efficient use of storage	The extent where the format represents data within the less possible storage cost [Redman96].
Format flexibility	The extent where the format is suitable to changes in user needs and the recording medium is not dramatically affected. The metric considers the number of format changes within the same recording medium (A), the number of format changes but with other recording medium (B).
Format Precision	When the set of symbolic representations are sufficiently precise to distinguish among elements in the domain that must be distinguished by the users, there are values correctly represented, values not represented (missing) and values that do not correspond with real world.
Interpretability	It is related to the format in which data are specified and to the clarity of data definitions. [Strong97].
Measurement	Is the act or process of quantitatively comparing results with requirements.
Portability	Is the extent where a format can be applied to a range of situations. The recording medium is important.
Price	Amount of money a user needs to pay for a query on a pay per query or pay per byte basis.
Relevance	Is the degree to which the provided data satisfies the users need.
Reliability	Data is reliable if it is considered as unbiased, good reputation and correct, complete and consistent.

Representation Consistency	It is whether physical instances of data are in accord with their format, the constraints are posed in terms of membership in the set of a symbolic representation [Redman96], or in terms of conformance to a format standard.
Reputation	Is the extent to which data are trusted or highly regarded in terms of their source or contents.
Response Time	Is the delay between the user request and the reception of the complete response from the Information System.
Timeliness	Is the extent to which the age of data is appropriate for the task at hand [Ballou98-2], and is computed in terms of currency and volatility.
Unbiased	Is the degree by which data is objective, and impartial.
Understandability	Is the degree to which the data can be easily comprehended by the user. Understandability measures how well a source presents its data.
Uniqueness	The extent where an entity from the real world is represented once.
Usability	Is the extent to which data are used for the task at a hand with acceptable effort. In other words if data come from a high reputed source, it is relevant to the task, it can be interpreted and understandable, and it provides benefit on the performance of the job. Usability is divided in usefulness and easy to use.
Usefulness	It is the degree where using data provides benefit on the performance on the job, in other words the extent to which the user believes data would be useful for the task at a hand.
Value Added	It is stated in terms of how easy is to get the task complete named as effectiveness; how long could the task take known as efficiency; and the personal satisfaction obtained from using data.
Value Consistency	Is the extent to which the values are the same for overlapping entities and attributes. Data are consistent with respect to a set of constraints if they satisfy all constraints in the set. [Motro98],

	[Redman96], and [Jarke98].
Verifiability	The degree to which data can be checked for correctness.
Volatility	Interval of time where data remains valid on the system, and it is related to the update frequency [Wang93], [Ballou98].

Appendix B TPC Benchmarks

We have implemented the benchmark datasets provided by the Transaction Processing Performance Council (TPC) for testing and experimenting purposes. The TPC defines transaction processing and database benchmarks to involve the measurement and evaluation of computer functions and operations. However, we have utilised the queries and transactions involved in their proposed environments to achieve our objectives.

This Appendix is aimed to present the main characteristics of TPC Benchmark™ H (TPC-H) and TPC Benchmark™ C (TPC-C). See [TPC] for further details.

B.1 TPC-H Benchmark

The TPC Benchmark™ H (TPC-H) is a decision support benchmark. It consists of a suite of business oriented ad-hoc queries and concurrent data modifications. The queries and the database structure have been chosen to have broad industry-wide relevance while maintaining a sufficient degree of ease of implementation.

TPC Benchmark™ H is comprised of a set of business queries designed to exercise system functionalities in a manner representative of complex business analysis applications. These queries have been given a realistic context, portraying the activity of a wholesale supplier.

TPC-H does not represent the activity of any particular business segment, but rather any industry which must manage, sell, or distribute a product worldwide (e.g., car rental, food distribution, parts, suppliers, etc.).

The selected queries provide answers to the following classes of business analysis:

- Pricing and promotions
- Supply and demand management
- Profit and revenue management
- Customer satisfaction study
- Market share study

- Shipping management.

B.1.1 Entity-Relationship Diagram of the TPC-H's business environment.

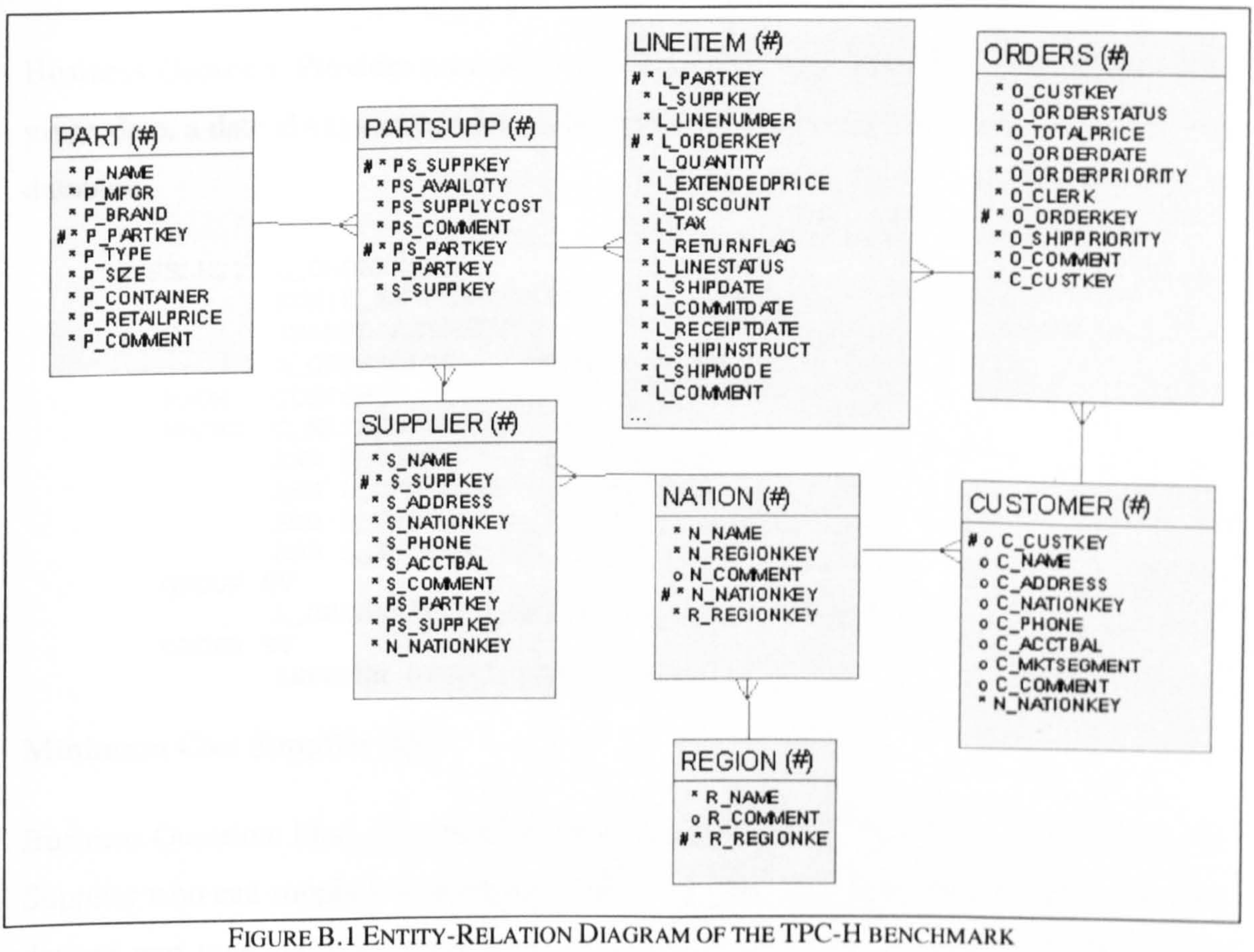


FIGURE B.1 ENTITY-RELATION DIAGRAM OF THE TPC-H BENCHMARK

B.1.2 The TPC-H Queries

Pricing Summary Report (Q1)

Business Question: Provides a summary pricing report for all Lineitems shipped as of a given date, a date always within 60 - 120 days of the greatest ship date contained in the database.

```
        SELECT  L_ORDERKEY,
                SUM(L_EXTENDEDPRICE*(1-L_DISCOUNT)(DECIMAL(18,2))
                (NAMED REVENUE),
                O_ORDERDATE, O_SHIPPRIORITY
        FROM    CUSTOMER, ORDERTBL, LINEITEM
        WHERE   C_MKTSEGMENT  = 'BUILDING'
                AND C_CUSTKEY = O_CUSTKEY
                AND L_ORDERKEY = O_ORDERKEY
                AND O_ORDERDATE < '1995-03-15'
                AND L_SHIPDATE > '1995-03-15'
        GROUP BY
                L_ORDERKEY, O_ORDERDATE, O_SHIPPRIORITY
        ORDER BY
                REVENUE DESC, O_ORDERDATE;
```

Minimum Cost Supplier (Q2)

Business Question: Find, in a given Region, for each Part of a certain type and size, the Supplier who can supply it at minimum cost. If several suppliers in that region offer the desired part type and size at the same (minimum) cost, the query lists the Parts from Suppliers with the 100 highest account balances.

```
SELECT  S_ACCTBAL, S_NAME, N_NAME, P_PARTKEY, P_MFGR,
        S_ADDRESS, S_PHONE, S_COMMENT
 FROM   PARTTBL, SUPPLIER, PARTSUPP, NATION, REGION
 WHERE  P_PARTKEY = PS_PARTKEY
        AND S_SUPPKEY = PS_SUPPKEY
        AND P_SIZE = 15
        AND P_TYPE LIKE '%BRASS'
        AND S_NATIONKEY = N_NATIONKEY
        AND N_REGIONKEY = R_REGIONKEY
        AND R_NAME = 'EUROPE'
        AND PS_SUPPLYCOST =
                (SELECT MIN(PS_SUPPLYCOST)
                FROM    PARTSUPP, SUPPLIER, NATION, REGION
                WHERE   P_PARTKEY = PS_PARTKEY
                        AND S_SUPPKEY = PS_SUPPKEY
                        AND S_NATIONKEY = N_NATIONKEY
                        AND N_REGIONKEY = R_REGIONKEY
                        AND R_NAME = 'EUROPE'
                )
ORDER BY
        S_ACCTBAL DESC, N_NAME, S_NAME, P_PARTKEY;
```

Shipping Priority (Q3)

Business Question: Retrieves the shipping priority and potential revenue of Orders having the largest revenue among those that had not been shipped as of a given date. Orders are listed in decreasing order of revenue. If more than 10 unshipped Orders exist, only the 10 Orders with the largest revenue are listed.

```
SELECT  L_ORDERKEY,
        SUM(L_EXTENDEDPRICE*(1-L_DISCOUNT)(DECIMAL(18,2))
        (NAMED REVENUE),
        O_ORDERDATE, O_SHIPPRIORITY
FROM    CUSTOMER, ORDERTBL, LINEITEM
WHERE   C_MKTSEGMENT  = 'BUILDING'
        AND C_CUSTKEY = O_CUSTKEY
        AND L_ORDERKEY = O_ORDERKEY
        AND O_ORDERDATE < '1995-03-15'
        AND L_SHIPDATE > '1995-03-15'
GROUP BY
        L_ORDERKEY, O_ORDERDATE, O_SHIPPRIORITY
ORDER BY
        REVENUE DESC, O_ORDERDATE;
```

Order Priority Checking (Q4)

Business Question: Counts the number of Orders ordered in a given quarter of a given year in which at least one Lineitem was received by the Customer later than its committed date. The query lists the count of such Orders for each order priority sorted in ascending priority order.

```
SELECT  O_ORDERPRIORITY, COUNT(*) (NAMED ORDER_COUNT)
 FROM   ORDERTBL
 WHERE  O_ORDERDATE >=  '1993-07-01'
        AND O_ORDERDATE < ADD_MONTHS('1993-07-01',3)
        AND EXISTS ( SELECT *
                     FROM    LINEITEM
                     WHERE   L_ORDERKEY = O_ORDERKEY
                             AND L_COMMITdATE < L_RECEIPtdATE
                     )
 GROUP BY
        O_ORDERPRIORITY
ORDER BY
        O_ORDERPRIORITY;
```

Local Supplier Volume (Q5)

Business Question: Lists for each Nation in a Region the revenue volume that resulted from Lineitem transactions in which the Customer ordering parts and the Supplier filling them were both within that Nation. The query considers only Parts ordered in a given year, displaying the Nations and revenue volume in descending order by revenue.

```
SELECT  N_NAME,
        SUM(L_EXTENDEDPRICE*(1-L_DISCOUNT) (DECIMAL(15,2)))
        (NAMED REVENUE)
 FROM   CUSTOMER, ORDERTBL, LINEITEM, SUPPLIER, NATION, REGION
 WHERE  C_CUSTKEY = O_CUSTKEY
        AND O_ORDERKEY = L_ORDERKEY
        AND L_SUPPKEY = S_SUPPKEY
        AND C_NATIONKEY = S_NATIONKEY
        AND S_NATIONKEY = N_NATIONKEY
        AND N_REGIONKEY = R_REGIONKEY
        AND R_NAME = 'ASIA'
        AND O_ORDERDATE >= '1994-01-01'
        AND O_ORDERDATE <  ADD_MONTHS('1994-01-01',12)
 GROUP BY
        N_NAME
 ORDER BY
        REVENUE DESC;
```

Forecasting Revenue Change (Q6)

Business Question: Considers all the Lineitems shipped in a given year with discounts between DISCOUNT - 0.01 and DISCOUNT + 0.01. The query lists the amount by which the total revenue would have increased if these discounts had been eliminated for Lineitems with L_QUANTITY less than an inputted quantity.

```
SELECT  SUM(L_EXTENDEDPRICE*L_DISCOUNT (DECIMAL(15,2)))
        (NAMED REVENUE)
 FROM   LINEITEM
 WHERE  L_SHIPDATE >= '1994-01-01'
        AND L_SHIPDATE <  ADD_MONTHS('1994-01-01',12)
        AND L_DISCOUNT BETWEEN .06 - 0.01 AND .06 + 0.01
        AND L_QUANTITY < 24;
```

Volume Shipping Query (Q7)

Finds, for two given Nations, the gross discounted revenues derived from Line Items in which parts were shipped from a Supplier in either Nation to a Customer in the other Nation during a period of time, commonly a year. Two Nations are given as input

parameters. This query is a 6-table join that requires a small table, Nation, to be aliased and processed as though it were two distinct look-up tables.

```
SELECT  N1.N_NAME, N2.N_NAME,
        EXtrACT(YEAR FROM L_SHIPDATE)(NAMED "YEAR"),
        SUM(L_EXTENDEDPRICE * (1-L_DISCOUNT) (DECIMAL(15,2)))
        (NAMED REVENUE)
 FROM   SUPPLIER, LINEITEM, ORDERTBL, CUSTOMER,
        NATION N1, NATION N2
 WHERE  S_SUPPKEY = L_SUPPKEY
        AND O_ORDERKEY = L_ORDERKEY
        AND C_CUSTKEY = O_CUSTKEY
        AND S_NATIONKEY = N1.N_NATIONKEY
        AND C_NATIONKEY = N2.N_NATIONKEY
        AND ((N1.N_NAME = 'FRANCE' AND N2.N_NAME = 'GERMANY')
        OR      (N1.N_NAME = 'GERMANY' AND N2.N_NAME = 'FRANCE'))
        AND L_SHIPDATE BETWEEN '1995-01-01' AND '1996-12-31'
GROUP BY
        N1.N_NAME, N2.N_NAME, "YEAR"
ORDER BY
        N1.N_NAME, N2.N_NAME, "YEAR";
```

National Market Share query (Q8)

The market share for a given Nation within a given Region is defined as the fraction of the revenue from the products of a specified type in that Region that was supplied by Suppliers from the given Nation. The query determines this for 2 years.

```
SELECT  EXtrACT(YEAR FROM O_ORDERDATE)
        (NAMED "YEAR"),
        SUM(CASE WHEN N2.N_NAME = 'BRAZIL'
                THEN  L_EXTENDEDPRICE*(1-L_DISCOUNT)
                ELSE 0
                END) / SUM(L_EXTENDEDPRICE*(1-L_DISCOUNT))
        (DECIMAL(15,2))
        (NAMED MKT_SHARE)
 FROM   PARTTBL, SUPPLIER, LINEITEM, ORDERTBL, CUSTOMER,
        NATION N1, NATION N2, REGION
 WHERE  P_PARTKEY = L_PARTKEY
        AND S_SUPPKEY = L_SUPPKEY
        AND L_ORDERKEY = O_ORDERKEY
        AND O_CUSTKEY = C_CUSTKEY
        AND C_NATIONKEY = N1.N_NATIONKEY
        AND N1.N_REGIONKEY = R_REGIONKEY
        AND R_NAME = 'AMERICA'
        AND S_NATIONKEY = N2.N_NATIONKEY
        AND O_ORDERDATE BETWEEN '1995-01-01' AND '1996-12-31'
        AND P_TYPE = 'ECONOMY ANODIZED STEEL'
GROUP BY
        "YEAR"
ORDER BY
        "YEAR";
```

Product Type Profit Measure (Q9)

Finds, for each nation an each year, the profit for all parts ordered in that year which contain a specific substring in their part names and which were filled by the Supplier in that nation.

```
SELECT  N_NAME
        ,EXtrACT( YEAR FROM O_ORDERDATE) (NAMED "YEAR")
        ,SUM((L_EXTENDEDPRICE*(1-L_DISCOUNT) -PS_SUPPLYCOST*L_QUANTITY)
        (DECIMAL(15,2)))
        (NAMED SUM_PROFIT)
FROM    PARTTBL, SUPPLIER, LINEITEM, PARTSUPP, ORDERTBL, NATION
WHERE   S_SUPPKEY = L_SUPPKEY
        AND PS_SUPPKEY = L_SUPPKEY
        AND PS_PARTKEY = L_PARTKEY
        AND P_PARTKEY = L_PARTKEY
        AND O_ORDERKEY = L_ORDERKEY
        AND S_NATIONKEY = N_NATIONKEY
        AND P_NAME LIKE '%green%'
GROUP BY
        N_NAME, "YEAR"
ORDER BY
        N_NAME, "YEAR" DESC;
```

B.2 TPC-C Benchmark

TPC Benchmark™ C is comprised of a set of basic operations designed to exercise system functionalities in a manner representative of complex OLTP application environments. These basic operations have been given a life-like context, portraying the activity of a wholesale supplier, to help users relate intuitively to the components of the benchmark. The workload is centred around the activity of processing orders and provides a logical database design, which can be distributed without structural changes to transactions.

TPC-C does not represent the activity of any particular business segment, but rather any industry, which must manage, sell, or distribute a product or service (e.g., car rental, food distribution, parts supplier, etc.). TPC-C does not attempt to be a model of how to build an actual application.

The purpose of a benchmark is to reduce the diversity of operations found in a production application, while retaining the application's essential performance characteristics, namely the level of system utilization and the complexity of operations. A large number of functions have to be performed to manage a production order entry system. Many of these functions are not of primary interest for performance analysis, since they are proportionally small in terms of system resource utilization or in terms of frequency of execution. Although these functions are vital for a production system, they merely create excessive diversity in the context of a standard benchmark and have been omitted in TPC-C.

B.2.1 Entity-Relationship Diagram of the TPC-C's business environment.

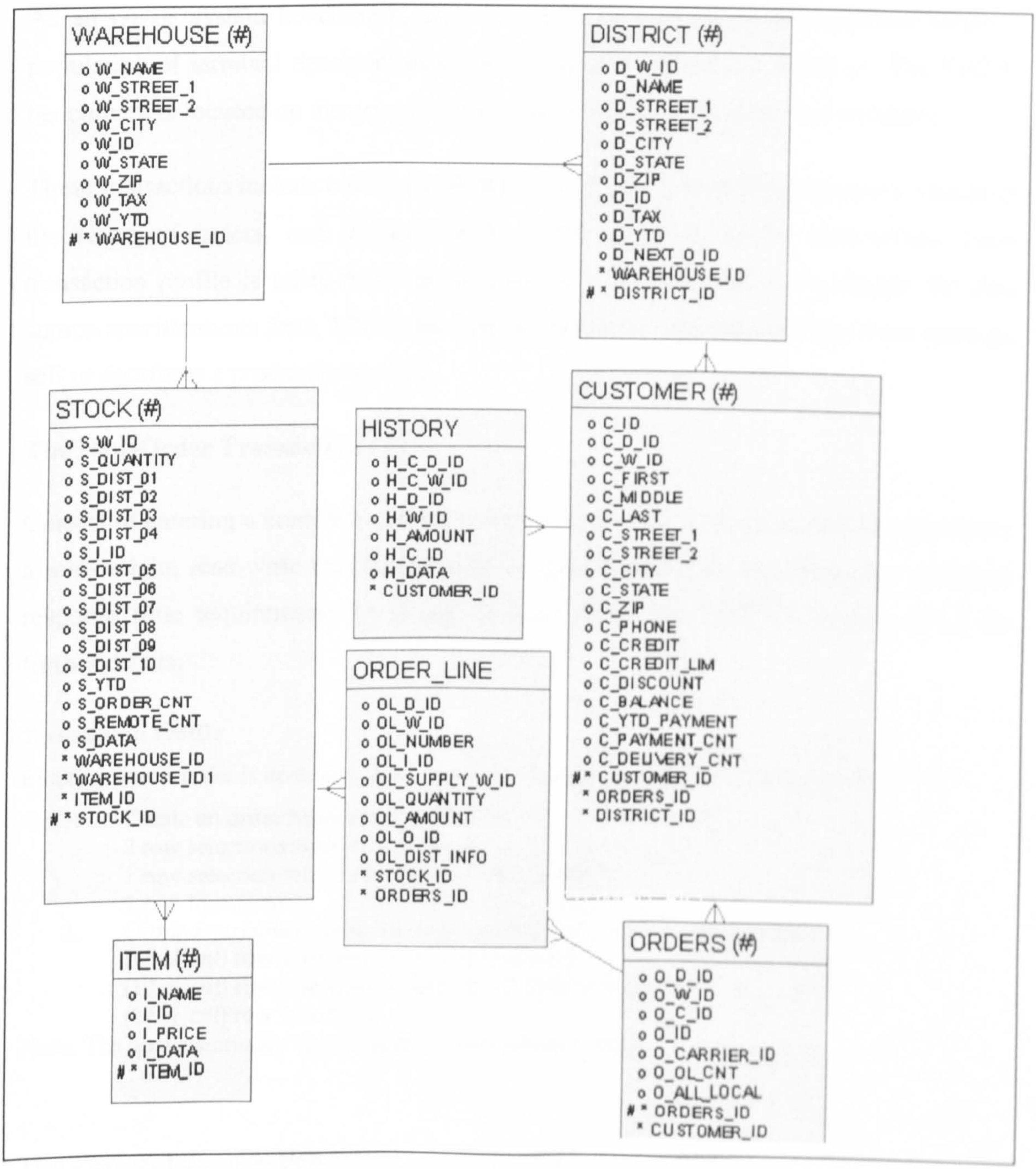


TABLE B.2 ENTITY RELATION DIAGRAM OF THE TPC-C BENCHMARK

B.2.2 Transaction Processing Systems of TPC-C

Aimed to process transactions in order to update records and generate brief reports.

As an OLTP system benchmark, TPC-C simulates a complete environment where a population of terminal operators executes transactions against a database. The TPC-C benchmark is focused on the principal transactions of an order-entry environment.

These transactions include entering and delivering orders, recording payments, checking the status of orders, and monitoring the level of stock at the warehouses. Each transaction profile is taken from the original text [TPC-C]. We have chosen the data source specifications from TPC-C because it represents any industry that must manage, sell or distribute a product or service.

The New-Order Transaction (T1)

Consist of entering a complete order through a single database transaction. It represents a mid-weight, read-write transaction with a high frequency of execution and stringent response time requirements to satisfy on-line users. See [TPC-C] section 2.4.2 for further details.

Transaction Profile

Entering a new order is done in a single database transaction with the following steps:

1. Create an order header, comprised of:
   2 row selections with data retrieval,
   1 row selection with data retrieval and update,
   2 row insertions.
2. Order a variable number of items (average *ol_cnt* = 10), comprised of:
   (1 * *ol_cnt*) row selections with data retrieval,
   (1 * *ol_cnt*) row selections with data retrieval and update,
   (1 * *ol_cnt*) row insertions.

Note: The above summary is provided for information only.

The Payment Transaction (T2)

The Payment business transaction updates the customer's balance and reflects the payment on the district and warehouse sales statistics. It represents a light-weight, read-write transaction with a high frequency of execution and stringent response time requirements to satisfy on-line users. In addition, this transaction includes non-primary

key access to the CUSTOMER table, because the customer might be selected on customer last name. See [TPC-C] section 2.5.2 for further details

Transaction Profile

The Payment transaction enters a customer's payment with a single database transaction and is comprised of:

Case 1, the customer is selected based on customer number:
3 row selections with data retrieval and update,
1 row insertion.
Case 2, the customer is selected based on customer last name:
2 row selections (on average) with data retrieval,
3 row selections with data retrieval and update,
1 row insertion.
Note: The above summary is provided for information only.

The Order-Status Transaction (T3)

The Order-Status business transaction queries the status of a customer's last order. It represents a mid-weight read-only database transaction with a low frequency of execution and response time requirement to satisfy on-line users. In addition, this table includes non-primary key access to the CUSTOMER table. See [TPC-C] section 2.6.2 for further details

Transaction Profile

Querying for the status of an order is done in a single database transaction with the following steps:

1. Find the customer and his/her last order, comprised of:
   Case 1, the customer is selected based on customer number:
   2 row selections with data retrieval.
   Case 2, the customer is selected based on customer last name:
   4 row selections (on average) with data retrieval.
2. Check status (delivery date) of each item on the order (average items-per-order = 10), comprised of:
   (1 * items-per-order) row selections with data retrieval.

Note: The above summary is provided for information only.

The Delivery-Transaction (batch processed) (T4)

The Delivery business transaction consists of processing a batch of 10 new (not yet

delivered) orders. Each order is processed (delivered) in full within the scope of a read-write database transaction. The number of orders delivered as a group (or batched) within the same database transaction is implementation specific. The business transaction, comprised of one or more (up to 10) database transactions, has a low frequency of execution and must complete within a relaxed response time requirement.

The Delivery transaction is intended to be executed in deferred mode through a queuing mechanism, rather than interactively, with terminal response indicating transaction completion. The result of the deferred execution is recorded into a result file. See [TPC-C] section 2.7.4 for further details

Transaction Profile

The deferred execution of the Delivery transaction delivers one outstanding order (average items-per-order = 10) for each one of the 10 districts of the selected warehouse using one or more (up to 10) database transactions. Delivering each order is done in the following steps:

1. Process the order, comprised of:
   1 row selection with data retrieval,
   (1 + items-per-order) row selections with data retrieval and update.
2. Update the customer's balance, comprised of:
   1 row selections with data update.
3. Remove the order from the new-order list, comprised of:
   1 row deletion.

Comment: This business transaction can be done within a single database transaction or broken down into up to 10 database transactions to allow the test sponsor the flexibility to implement the business transaction with the most efficient number of database transactions.

Note: The above summary is provided for information only.

The Stock-level Transaction (T5)

The Stock-Level business transaction determines the number of recently sold items that have a stock level below a specified threshold. It represents a heavy read-only database transaction with a low frequency of execution, a relaxed response time requirement, and relaxed consistency requirements. See [TPC-C] section 2.8.2 for further details

Transaction Profile

Examining the level of stock for items on the last 20 orders is done in one or more database transactions with the following steps:

1. Examine the next available order number, comprised of:
   1 row selection with data retrieval.
2. Examine all items on the last 20 orders (average items-per-order = 10) for the district, comprised of:
   (20 * items-per-order) row selections with data retrieval.
3. Examine, for each distinct item selected, if the level of stock available at the home warehouse is below the threshold, comprised of:
   At most (20 * items-per-order) row selections with data retrieval.

Note: The above summary is provided for information only.

Appendix C Statistical Tables

This Appendix presents a decision table for selecting the appropriate statistical procedure, and the statistical tables utilized for the experiments.

C.1 Table for Inferential Statistical Tests with Ordinal/Rank-Order Data

	Type of Data			
Goal	Measurement (from Gaussian Population)	Rank, Score, or Measurement (from Non- Gaussian Population)	Binomial (Two Possible Outcomes)	Survival Time
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample *t* test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired *t* test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired *t* test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*

TABLE C.1. SELECTING A STATISTICAL TEST FROM [MOTULSKY95]

C.2 Table of critical values for the Wilcoxon test

One Tailed Significance Levels				
	0.025	0.01	0.005	0.001
Two Tailed Significance Levels				
N	0.05	0.02	0.01	0.002
6	0	-	-	
7	2	0	-	
8	3	1	0	
9	5	3	1	
10	8	5	3	0
11	10	7	5	1
12	13	9	7	2
13	17	12	9	4
14	21	19	15	6
15	25	20	16	8
16	29	24	20	11
17	34	28	23	14
18	40	33	28	18
19	46	38	32	21
20	52	43	38	26
21	58	49	43	30
22	65	56	49	35
23	73	62	55	40
24	81	69	61	45
25	89	77	68	51
26	98	84	75	58
27	107	92	83	64
28	116	101	91	71
29	126	110	100	79
30	137	120	109	86
31	147	130	118	94
32	159	140	128	103
33	170	151	138	112
34	182	162	148	121

TABLE C.2 TABLE OF THE CRITICAL VALUES FOR THE WILCOXON TEST

Method: Compare the obtained value of Wilcoxon's test statistic to the critical value in the table considering N as the number of subjects. The obtained value is statistically significant if it is equal to or smaller than the value in the table. For instance, suppose an obtained value is 71, with 23 subjects. The critical value in the table is 73: so the obtained value is smaller than this. Therefore the conclusion would be that the difference between the two conditions in the experiment was unlikely to occur by chance ($p<.05$ two-tailed test, or $p<.025$, one-tailed test).

C.3 Table of Critical Values for Spearman's Rho (r)

	One Tailed level of significance		
	0.025	0.01	0.005
N	Two Tailed level of significance		
	0.05	0.02	0.01
5	1	1	-
6	0.886	0.943	1
7	0.786	0.893	0.929
8	0.738	0.833	0.881
9	0.683	0.783	0.833
10	0.648	0.746	0.794
12	0.591	0.712	0.777
14	0.544	0.645	0.715
16	0.506	0.601	0.665
18	0.475	0.564	0.625
20	0.45	0.534	0.591
22	0.428	0.508	0.562
24	0.409	0.485	0.537
26	0.392	0.465	0.515
28	0.377	0.448	0.496
30	0.364	0.432	0.478
35	0.335	0.394	0.433

TABLE C.3 TABLE OF CRITICAL VALUES FOR NON-PARAMETRIC SPEARMAN CORRELATION.

Method: Consider the number of pairs of scores (N) that were used in the experiment, and then compare the obtained value of r to the value in the corresponding column.

For instance: The r value of 0.3822, with 26 pairs of scores, is smaller than the critical value of r at the 0.05 level of significance (0.392). (i.e. it is highly significant). If your N is not in the table, use the next one down - e.g., for an N of 27, use the table values for 26.

C.4 Table of the Chi-Square Distribution

df \ *p*	.005	.01	.025	.05	.10	.90	.95	.975	.99	.995
1	.00004	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
.6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.6	5.23	6.26	7.26	8.55	22.31	25	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

TABLE C.4 TABLE OF THE CHI-SQUARE DISTRIBUTION

The Friedman Test

Purpose: The Friedman test, also known as Friedman two-way analysis of variance, tests the null hypothesis that measures from k dependent samples come from the same population. It is based on the rationale that if the groups do not differ on the criterion variable, then the rankings of each subject will be random and there will be no difference in mean ranks between groups on the criterion variable.

Calculation. The Friedman test statistic is distributed approximately as chi-square, with (k - 1) degrees of freedom, where k is the number of groups in the criterion variable, from i = 1 to k. Let n be the number of subjects and let Ti be the sum of ranks for each group. Friedman chi-square is then computed by this formula:

$$Fr = \frac{12}{nk(k+1)}\left[\sum_{j=1}^{k} Ti^2\right] - 3n(k+1)$$

Method: In order to reject the null hypothesis, the computed value Fr must be equal to or greater than the tale critical chi-square value at the specified level of significance. For the appropriate degrees of freedom, the tabled Fr .95 value (which is the chi-square value at the 96th percentile) and the tabled Fr.99 value (which is the chi-square value at the 99th percentile) are employed as the 0.05 and 0.01 critical values for evaluating a nondirectional alternative hypothesis. The number of degrees of freedom are computed as df = k-1.

Appendix D User Stereotypes

Analysis of Data Quality properties, types of user and Information Systems

The material presented in this Appendix still on working process as part of future work.

D.1 Introduction

We have proposed to develop a number of user stereotypes to support different contexts and help novice users to specify query context during the analysis of data quality in the DQM prototype. The identification of user stereotypes still on working process, therefore such stereotypes are part of future work. However, this Appendix discusses the status of our research.

User stereotypes will determine which quality criteria to use, and the inter-dependencies between them, based on the application domain and the type of users. This work is a complement to the framework for data integration considering Data Quality in [Angeles04], and the quality interdependencies have been published in [Angeles05].

Different levels of customer satisfaction can be deduced, from the proper accomplishment of their requirements as is indicated in [Kano84].

User preferences may vary depending on the information system. On one hand, data consumers of a DSS might prefer some data against other because of reputation of data producers, the credibility and relevance of data for the task at a hand, or the level of satisfaction they have on making strategic decision effectively from using reliable data. On the other hand, data consumers of operational systems might be more interested in timeliness, availability and accessibility of data for an effective transaction processing.

User preferences may vary depending on their experience as product of their knowledge. This Appendix will discuss the types of users how they may prefer some quality properties depending on their role. The identified dependency among quality properties for a subjective measurement and after that, the types of Information

Systems, and their relevant quality properties. Finally, we present the our approximation of the user stereotypes.

D.2 Types of users

As we have mentioned before, the role and experience of users determine the context of the query and therefore the query outcomes. According with Lee and Strong in [Lee04], the responses from three roles within data production process determine data quality because of their knowledge.

The three roles mentioned were data collector, data custodian, and data consumer. The Lee et al research was oriented to determine the causes of poor data quality during the data life cycle and how the knowledge of the participant users reflects the quality of data. We take from this research the hypothesis that knowledge is important on the quality of data. In our research, we are more interested in identifying which are the relevant quality properties from each type user perspective in order to query data not to produce data.

We consider that the quality properties they would be interested in during the execution of a query might be much related to their work role. The following types of users were defined in [Lee04] as data collectors (people or groups who generate information); data custodians (people or groups who manage computing resources for storing and processing data), and data consumers (people or group who use data, which may be involve in retrieval of data, additional data aggregation and integration. We consider such type of users for the identification of our user stereotypes. Lee considered the following data quality properties: accessibility, relevance, timeliness, completeness, and accuracy.

Data quality is associated with know why from data collectors with accuracy, accessibility, relevance, completeness and timeliness.

DQ is associated with data custodian with know what and accuracy, completeness and timeliness.

DQ is associated with data consumer with know how and know why and relevance.

If we just consider the type of users with the data quality properties, that we have identified in our Reference Model relative to the entire data quality life cycle stated in

Chapter 3, we could say that collectors are related to the data value level properties situated at the product base level. As the work role of the custodians is responsible for the storage and maintenance of data they are directly concerned with data representation at the design level. Data consumers are therefore interested in the quality properties proper to data context at the external focus for information retrieval and utilisation. Our first approximation for the relation between quality properties with types of users is as shown as follows:

Data Collector
- Accuracy
- Completeness
- Currency
- Timeliness
- Uniqueness
- Value consistency
- Volatility

However, Lee in [Lee04] considers accessibility and relevance as part of quality properties relative to data collectors.

Data Custodian
- Ability to represent nulls
- Appropriateness
- Appropriateness of Format
- Concise Representation
- Efficient use of storage media
- Format Flexibility
- Format Precision
- Representation Consistency
- Understanding
- Value Consistency

In the case of data custodians, Lee et al identified accuracy, completeness and timeliness.

Data Consumer	Accessibility Amount of data Availability Believability Cost Ease to use Effectiveness Efficiency Interpretability Price Relevance Reliability Reputation Response Time Unbiased Understand ability Usefulness Value Added Verifiability

In the case of data consumer, Lee et al identified relevance.

As we can observe, this classification is not as straightforward as we could think. "All modes of knowledge combined about all processes held by all roles, contribute to the overall quality" [Lee04]. Besides, users may have different perceptions depending on their experience, information system, etc. For instance, if we consider volatility as "the rate of change of the real work" it maybe related to the data collector, because he is on charge to reflect that change. However, if we consider volatility as the update frequency then it changes according with the application domain. Another example is the event when data consumer assesses interpretability, but the data custodian is the originator.

D.3 Data Quality Interdependencies

The measurement of a quality criterion might be part of the measurement of an aggregate one. The quality dimensions, which measurements derive from primary criteria, are identified as secondary quality properties. However, we have not established or tested any kind of correlation among them. We have identified some relationships between these quality properties based on their definitions from previous research. In this section, the secondary quality criteria definitions and their relationship with primary criteria are as follows:

D.3.1 Primary Quality Criteria

From the Data Quality Reference Model, we have identified a number of criteria which measurement *does not depend* on other quality criteria, namely Primary Quality criteria [Angeles05], (See Table D.2).

Accuracy	Format Precision
Currency	Format Flexibility
Efficient use of storage	*Volatility*
Response time	Representational *Consistency*
Availability	Concise Representation
Amount of data	Appropriateness of Format
Unbiased data	Uniqueness

TABLE D.1 PRIMARY QUALITY CRITERIA

D.3.2 Secondary quality Criteria

This section presents the secondary quality criteria and their dependency among data quality criteria. These dependencies can help in the measurement of such quality properties, most of them by subjective methods.

The interpretability is related to the format in which data are specified an to the clarity of data definitions in [Strong97]. Thus, it depends on several factors: If there is any change on user needs, its representation should not be affected, this can be possible with a flexible format; The data value shall be presented consistently through the application or applications; and that the format is sufficient to represent what is needed and in the proper manner.

Reputation is the extent to which data are trusted or highly regarded in terms of their source or content. Three factors shall be considered at measuring time: reputation of data should be determined by its overall quality. If authors of data provide inaccurate data then they are unreliable and shall be therefore decreased. Commonly reputation might be increased if authors have enough experience gained across the time. If data owners produce accurate data consistently, modify data as soon as possible when mistakes are found, and they in turn recommend authors of quality data.

Accessibility is the extent to which data is accessible in terms of availability and security and cost. On one hand, data might be available but inaccessible for security purposes. On another hand, data might be available but expensive.

Data is *credible* as true if it is correct, complete, and consistent.

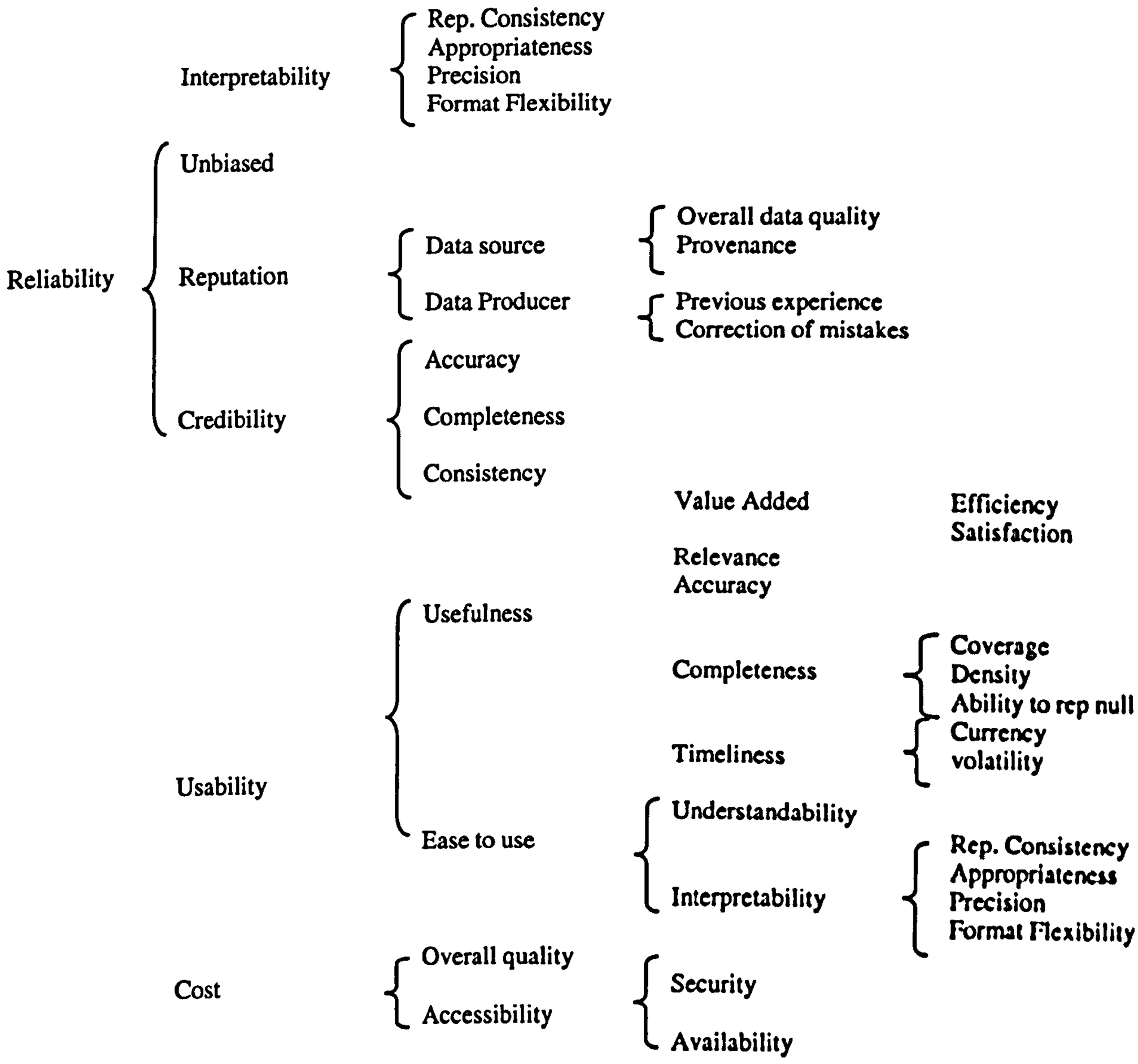
Usability is the extent to which data are used for the task at a hand with acceptable effort. In other words if data come from a high reputed source, it is relevant to the task, it can be interpreted and understandable, and it provides benefit on the performance of the job. Usability is divided in usefulness and easy to use.

Usefulness is the degree where using data provides benefit on the performance on the job, in other words the extent to which the user believes data would be useful for the task at a hand. (value added, relevance, accuracy, completeness and timeliness).

Easy to use is the degree of effort user needs to apply to use data. As lesser effort as easier to use. This effort is in terms of understandability and interpretability as resources needed to achieve the expected goals. The personal factor in the usability lays on the reputation. Commonly user use determined data sources, due to the reputation of authors. The measurement of usability allow user to decide on the acceptance of data, and select a specific datum, data or data source among other alternatives.

Data is *reliable* if it is considered as unbiased, good reputation and correct, complete and consistent.

The *value added* is stated in terms of how easy is to get the task complete named as effectiveness; how long could the task take known as efficiency; and the personal satisfaction obtained from using data.



Other factor to consider is that data quality properties vary depending on the type of Information System.

D.4 Types of Information System

D.4.1 On-Line Transaction Processing (OLTP)

The purpose of these systems is to answer routine questions and track the flow of transactions through the organisation to help operational supervisors. Applications such as billing systems, cash deposits, withdraws, credit decisions, stock control systems, flight reservations, so production and purchasing systems, sales are typical operational systems.

Operational systems manage small quantity of data per transactions, Accuracy, response time, and currency, are important at this level of information systems. In consequence, data have high volatility due to very frequent updates. The record might not be complete, but they are fundamental data for the transaction that must be accurate.

For instance, in financial services such as cash machines concise representation, and representation consistency and format precision is important. However response time, and availability are more important along with security, because usually are systems with millions of customers.

All this involves the usability of data and preference from one cash account or credit account with respect to other plus cost of using and administering the service.

On one hand there are cases where service is expensive due to high cost of administering data, but their reputation is high an on the other hand service could be less expensive, but the company is recently established.

D.4.2 Management Information Systems and Decision Support Systems

Management Information Systems are mainly concerned with monitoring, controlling, decision-making and administrative activities. MIS usually take data from the transactional processing systems as internal source of information as from external data sources. MIS reports end to be used by middle management and operational supervisors.

Transactional data must be complete and correct in order to generate good summarize data, as their analysis implies high volume of data and routine reports, data must be timely before the analysis.

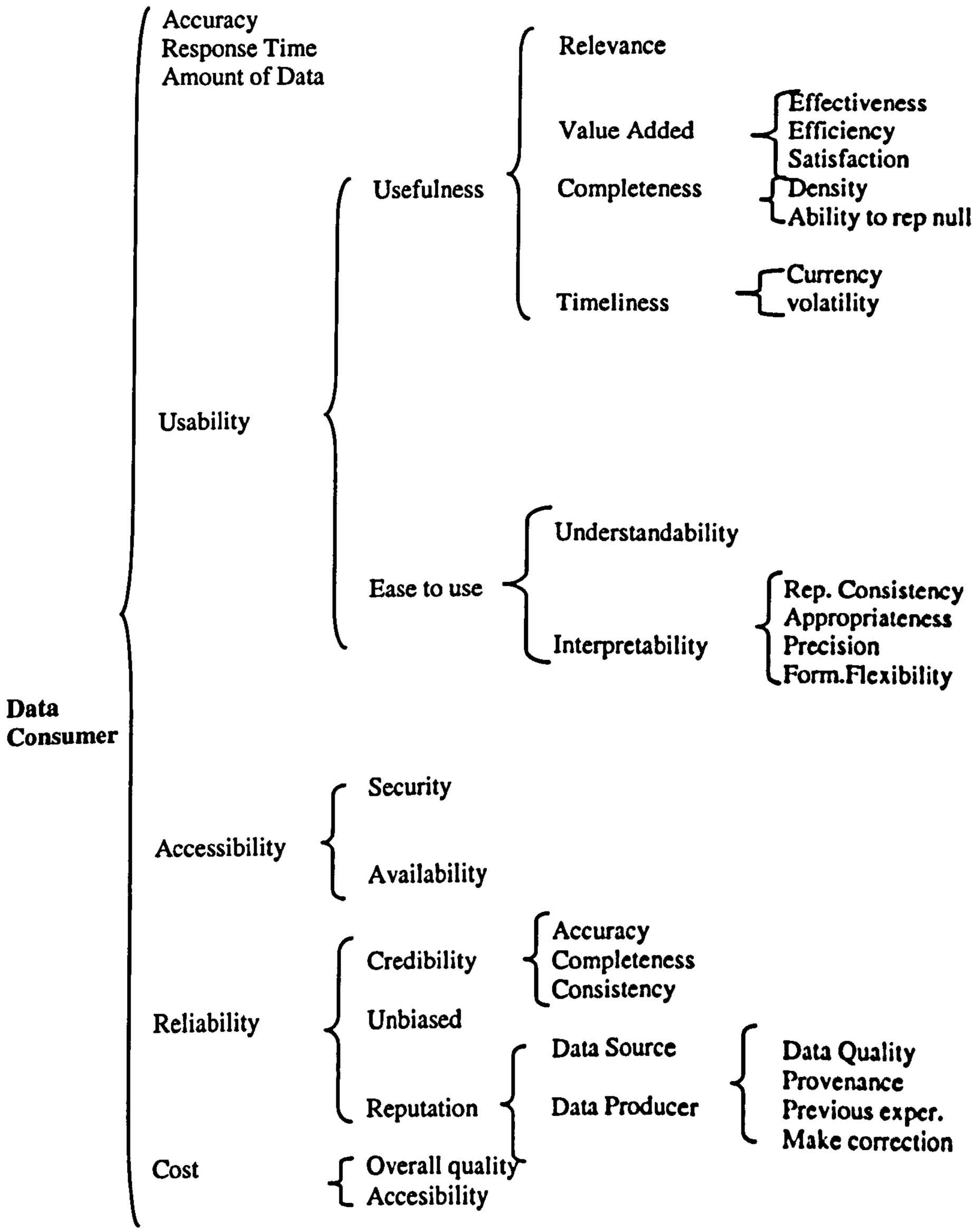
Decision Support Systems (DSS): Help professional and staff managers make decisions. DSS comprises techniques and tools to help gather relevant information and analyse the options and alternatives.

In a Data Warehouse, the quality criteria vary depending on the data source, for example for look up tables there will be low volatility, but accessibility is important. In case of Fact tables, as they provide the sales detail, accuracy, uniqueness, and completeness are important because they would be directly reflected in the generation of aggregate data in the summarize tables.

An Executive Support System is designed to help a senior management tackle and address issues and long-term trends to make strategic decisions for the business. It gathers analyses and summarises aggregate, internal and external data to generate projections and responses to queries.

The main data quality problem on EIS relays on external data, so decisions depend on accuracy, timeliness, completeness and reliability of external data sources. Reputation, portability, relevance, and amount of data are quite important. Data consumers require friendly and usable tools in order to deal just with making decisions rather the IS per se. Possible inconsistencies might be derived from different data sources so making decisions regarding which external data source to trust is an issue. Response time however, is not of great relevance when the analysis in on long-term trends.

Data Consumer with DSS



Data Custodian in OLTP